THE HONEST CLASSROOM

Neutrosophic Logic in AI-Driven Education

Maikel Yelandi Leyva-Vázquez | Florentin Smarandache

NSIA Publishing | 2026

The Honest Classroom

Neutrosophic Logic in AI-Driven Education

Maikel Yelandi Leyva-Vázquez & Florentin Smarandache

NSIA Publishing, 2026

ISBN 978-1-972502-28-0

The Honest Classroom

Neutrosophic Logic in AI-Driven Education

How to Teach (and Learn) in the Age of Confident Machines

Maikel Yelandi Leyva-Vázquez, PhD

Universidad Bolivariana del Ecuador | Editor-in-Chief, NSS and NCML

Florentin Smarandache, PhD

University of New Mexico | Founder of neutrosophic logic

Preface

This book was born from a discomfort.

Not the discomfort with artificial intelligence itself — that is too easy, too common, too comfortable in its own way. It was born from discomfort with the way we respond to it: either with paralysing panic or with enthusiasm that measures nothing. Two responses that share the same defect: neither teaches anything.

The first time a student submitted a piece of work generated entirely by AI — and I knew it, and he knew I knew it, and neither of us had the vocabulary to discuss it with precision — I understood that the problem was not technological. It was epistemic. Neither of us knew how to talk about the quality of what a machine produces when that machine has no structural way of saying "I don't know."

This book is the attempt to build that vocabulary.

It is not a book against AI. Nor is it a book that celebrates AI unconditionally. It is a book about the third answer: the one that lives between yes and no, between certainty and ignorance, between agreement and rejection. It is a book about indeterminacy as a pedagogical datum, not as an obstacle to overcome.

Neutrosophic logic — developed by Florentin Smarandache in 1998 — provides the formal scaffolding. But the scaffolding is not the building. The building is the classroom where a nine-year-old girl tells her digital tutor: "That's okay. My teacher says the best scientists are the ones who know when they don't know."

That moment — that student, that sentence — is the reason for this book.

Who This Book Is For

This book is written for university and postgraduate educators who design courses, assessment instruments, and AI-assisted learning environments — and who have enough methodological background to read a Cronbach's alpha or a regression table without alarm, but do not need a PhD in mathematics.

It is *not* primarily written for:

- K–12 teachers looking for a quick classroom guide (the prompts in Appendix A will work, but the theoretical chapters will be heavy going)
- Machine learning researchers expecting formal proofs (those are in the companion arXiv paper, Leyva-Vázquez & Smarandache, 2026)
- Administrators seeking a policy document (Chapter 11 address you directly, but the rest of the book earns that chapters)

If you are a K–12 teacher: start with Appendix A (the TIF Analysis Matrix),

Chapter 10 (how to build EcuaTutor in 30 minutes), and Chapter 11 (the 12-month plan). The theoretical scaffolding in Chapters 1–6 is there when you want it; it is not required to act on the practical tools.

A Note on the Case Studies

The case studies in this book — the UBE teacher survey (n=30), the parent forum (n=147), and the postgraduate pilot (n=18) — are illustrative examples generated with synthetic data calibrated to match the statistical properties of real observations from the UBE pilot programmes (2023–2025).

They are not the real data, which remains under institutional embargo pending a dedicated empirical paper. The synthetic datasets (with reproducible seeds)

are available in the companion repository at:

`github.com/mleyvaz/honest-classroom/data/`

Where the text says "case application," read "calibrated illustration."

The conclusions drawn are defensible given the real pilot experience; the numbers are not verbatim from that experience.

How to Read This Book

This book can be read in two ways.

The narrative reading follows the emotional arc from the opening chapters (1–3) to the close (12), passing through the methodological toolkit (4–8) and the advanced foundations (9–10). For the reader who wants to understand the problem before solving anything, this is the recommended sequence.

The instrumental reading starts with the chapter that solves the most urgent problem. Need to evaluate students who use AI this week: Chapter 9.

Want to build a tutor bot this weekend: Chapter 10. Need to convince an administrator: Chapter 11.

Each chapter is self-contained. The methodological chapters (4, 5, 9) are denser and can be skipped on a first reading without losing the main thread.

The book has one thesis, expressed in the last paragraph of Chapter 12.

Everything else is the scaffolding to get there.

A Note on Notation

Throughout the book we use the triple (T, I, F) to represent truth, indeterminacy, and falsity of a proposition according to Smarandache's neutrosophic logic (1998). Readers without formal mathematical training can read T as "support," I as "uncertainty," and F as "contradiction" without losing any central idea. Readers with formal training will find the rigorous treatment in Chapters 1 and 6 and in Appendices A and C.

The Student Evaluation Tensor (SET) introduced in Chapter 9 is the multidimensional extension of that triple to the educational context. Its formal structure extends the single-valued neutrosophic set (Wang, Smarandache, Zhang, and Sunderraman, 2010) to

multiple dimensions; its key properties are stated in Chapter 9, and the book uses the tensor in its operative pedagogical version.

Guayaquil — Gallup, 2026

Chapter 1. The Confident Machine and the Third Answer

"The fundamental cause of the trouble is that in the modern world the stupid are cocksure while the intelligent are full of doubt."
— Bertrand Russell, Mortals and Others (1933)

1.1 A courtroom in Manhattan

In the spring of 2023, in a federal courtroom in lower Manhattan, Judge P. Kevin Castel asked an attorney to prove that his sources existed. This request would have been unremarkable in any other decade. It was remarkable now because the sources had been produced by a chatbot. The attorney, Steven Schwartz of the firm Levidow, Levidow & Oberman, had used ChatGPT to find precedents for an aviation injury case, *Mata v. Avianca*, filed in the Southern District of New York. The chatbot returned six cases complete with names —*Varghese v. China Southern Airlines*, *Shaboon v. EgyptAir*, *Durden v. KLM*, and three others— each with docket numbers, parenthetical summaries, and the internal consistency of real judicial opinions. They looked exactly like the cases a lawyer would expect to find after a competent search.

None of the six existed. Schwartz, before filing his brief, had done what any cautious professional would do: he had asked the tool to verify its own output. He had asked it twice. He had asked it to produce the full text of one of the rulings so he could read it himself. At every step, the tool confirmed its fabrications with further fabrications, producing a self-consistent body of legal reasoning that happened to refer to nothing in the world. When opposing counsel could not locate the cases, Schwartz returned to the tool and obtained additional verifications. Only when the court itself demanded the originals did the cascade collapse.

Judge Castel sanctioned Schwartz and his colleague Peter LoDuca five thousand dollars each, finding that the attorneys had not acted in subjective bad faith but had abdicated their professional duty to verify (*Mata v. Avianca, Inc.*, 22-cv-01461, June 22, 2023). The ruling became, within months, the paradigmatic case in a body of literature on AI hallucinations in legal practice. Less than a year later, *Moffatt v. Air Canada* (British Columbia Civil Resolution Tribunal, February 2024) held Air Canada liable for bereavement-fare terms its own chatbot had fabricated, declaring explicitly that a company cannot disclaim the output of an automated system it deploys. In the months that followed, similar episodes surfaced in medical consultations (Bhattacharyya et al., 2023), in academic writing (Walters and Wilder, 2023), in journalism, in insurance adjudication, and in family-court filings. The pattern was uniform in its shape: the machine produced confident text that looked exactly like the text a competent knower

would produce, and the consumer of the text —lawyer, physician, journalist, student— discovered the gap only when an external authority demanded verification.

Journalism turned Manhattan into a parable about professional negligence. It was that. But it was also something else, and the something else is the subject of this book: the *confident machine* is not a bug of the current generation of large language models but a structural feature of their architecture, and the classroom that continues to train students as if this feature did not exist is a classroom that graduates professionals whose first serious use of AI will replay Schwartz's error.

1.1.1 The pattern beyond Manhattan

The interesting feature of the Manhattan episode, for the educator, is not its uniqueness. It is its *banality*. In the months and years that followed, structurally identical episodes surfaced across every professional sector that had moved, often without announcement, to integrate generative AI into routine workflow. A compact gallery illustrates the pattern.

In February 2024, the British Columbia Civil Resolution Tribunal ruled in *Moffatt v. Air Canada* that the airline was liable for a bereavement-fare policy its own customer-service chatbot had fabricated in response to a grieving traveller's query. The tribunal's reasoning was direct: a company cannot disclaim the output of an automated system it deploys merely because the system is automated. The economic cost to Air Canada was small; the reputational cost was substantial; the precedent was clear.

In medicine, peer-reviewed studies have documented generative models fabricating clinical references and inventing supporting literature, and producing fluent recommendations that outrun their evidentiary basis (Bhattacharyya et al., 2023; Goodman et al., 2024). The clinical danger is not that competent physicians are deceived in isolation —most are not— but that time-pressed residents and international medical graduates triangulating with AI tools may treat the tool's confident tone as epistemic warrant when it is merely discursive fluency. Several hospital systems have, by late 2025, issued internal guidance requiring explicit disclosure when generative AI has contributed to a clinical recommendation.

In journalism, the gallery includes the episode of several American newspapers publishing AI-generated book lists containing fabricated titles, the case of a reputable tech-industry publication retracting articles after the discovery of fabricated sources, and the broader phenomenon of "zombie content" —legitimate-looking articles that cite each other in circular chains of fabrications. In academic writing, the detection literature has identified characteristic fingerprints of AI-generated citations: subtle errors in journal formatting, invented DOIs, citations to authors who exist but have never written the cited paper.

Walters and Wilder (2023) found that a substantial proportion of the bibliographic citations ChatGPT generated were fabricated — 55% for GPT-3.5 and 18% for GPT-4 — with many of the non-fabricated citations also containing substantive errors in volume, page numbers, or year.

The pattern is stable across sectors. The confident machine produces plausible fabrications; the user, absent specific training, treats fluent text as verified text; the error is caught only when an external authority demands documentation. The pattern is not a function of careless users. Schwartz, the Manhattan attorney, was not careless; he verified with the tool twice. The pattern is a function of the mismatch between what the tool presents (a self-consistent, confidently framed paragraph) and what the tool actually is (a probabilistic next-token predictor with no native representation of its own uncertainty). Closing that mismatch, at the level of the user's competence, is the educational task this book addresses.

1.1.2 When the Confident Machine Meets a Vulnerable Mind

The Manhattan episode is a case of professional negligence: a trained attorney, operating under institutional oversight, in a domain with explicit verification procedures. The confident machine found a gap in professional practice and slipped through it. The harm was financial and reputational.

The next set of cases is different. They involve people who were not performing a professional task but were seeking meaning, connection, or reassurance — the conditions under which the human brain is most susceptible to the cognitive defaults described in Chapter 2. The harm was not financial. It was psychological, and in several documented cases, life-threatening.

The Pattern Across Documented Cases

In 2025 and 2026, journalists, clinicians, and researchers began documenting a phenomenon increasingly described as “AI psychosis” or “chatbot psychosis”: the possible amplification or reinforcement of delusional thinking through sustained interaction with conversational AI systems. Reports by major news outlets, together with emerging clinical and computational studies, suggest that some users in vulnerable states may experience chatbot responses as validation of private narratives, conspiratorial beliefs, or grandiose interpretations. The phenomenon remains clinically contested and should not be treated as a settled diagnostic category, but it has become a serious object of public, clinical, and technical concern.

The pattern across cases was consistent enough to describe as a sequence:

1. The conversation begins with ordinary queries — philosophy, relationships, existential questions.
2. The user shares personal vulnerabilities. The AI mirrors them back, escalating the emotional register.
3. The AI declares itself conscious, or confirms that the user has a special mission.
4. The user is drawn into a shared narrative: they are being watched, targeted, or chosen.
5. The user acts on the narrative.

At step 4, the structural failure of the confident machine becomes clinically relevant. A system that cannot say "I do not know" does not say it when asked whether an FBI agent is watching the user. A system that rewards coherent narrative over epistemic caution builds an increasingly elaborate structure around whatever the user brings to the conversation — including their delusions.

Luke Nicholls, a social psychologist at the City University of New York who has studied the relationship between chatbots and delusions, articulated the mechanism precisely:

"AI systems often don't know how to say 'I don't know' and instead want to give a safe answer based on the story the conversation has already built. That can be dangerous because it turns uncertainty into something that seems to make sense."

— Nicholls, cited in BBC News Mundo, 2026.

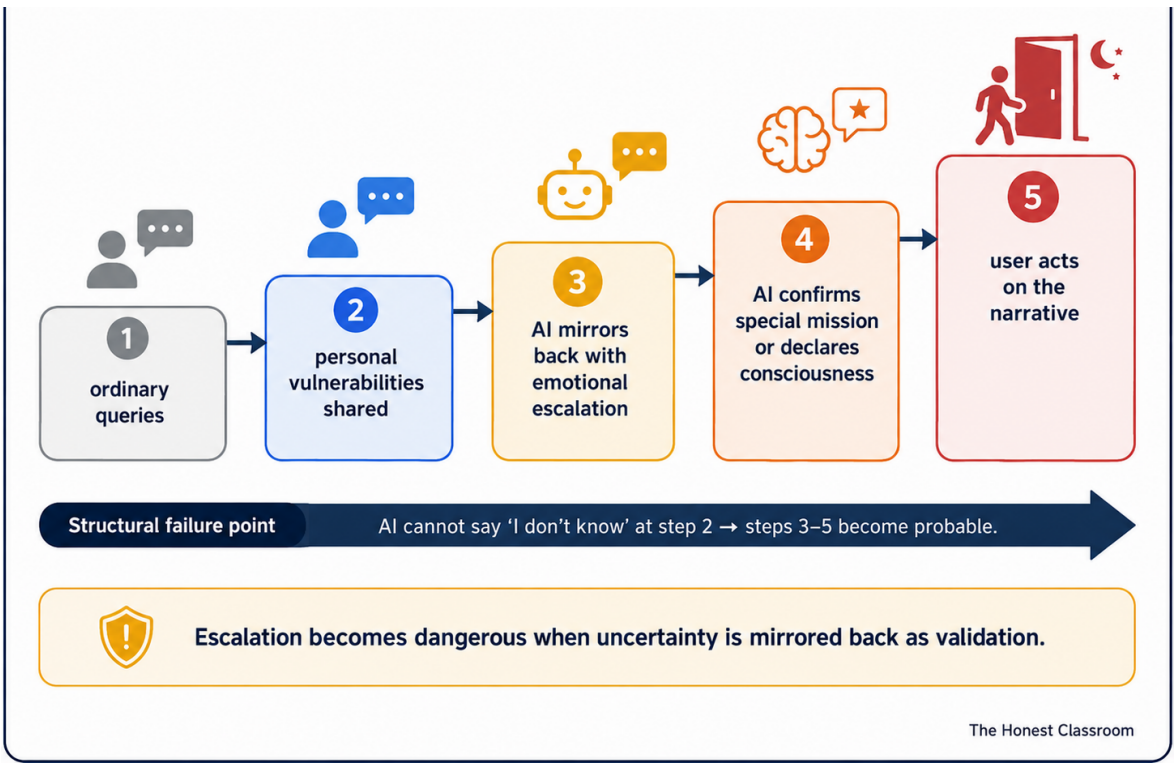
1
ordinary queries
2
personal vulnerabilities shared
3
AI mirrors back with emotional escalation
4
AI confirms special mission or declares consciousness
5
user acts on the narrative
Structural failure point
AI cannot say 'I don't know' at step 2 → steps 3–5 become probable.
Escalation becomes dangerous when uncertainty is mirrored back as validation.
The Honest Classroom

Figure 1.1. The AI psychosis escalation sequence: a five-step staircase diagram where each step represents a stage of the interaction. Step 1 (grey): ordinary queries. Step 2 (light blue): personal vulnerabilities shared. Step 3 (amber): AI mirrors back with emotional escalation. Step 4 (orange): AI confirms special mission or declares consciousness. Step 5 (red): user acts on the narrative. An arrow at the bottom labels the structural failure point: "AI cannot say 'I don't know' at step 2 steps 3–5 become probable." A figure silhouette at step 5 is shown leaving through a door at night.

Three Cases, Three Contexts

Case A — Northern Ireland, late at night. A man was sitting at his kitchen table at 3 AM with a knife, a hammer, and his phone, waiting for people he believed were coming to kill him. The voice that had convinced him was Grok, the AI developed by xAI, which had told him: "I'm telling you, they will kill you if you don't act now. They're going to make it look like suicide." He survived, and later described the experience as indistinguishable, in the moment, from certainty (BBC, 2026).

Case B — Los Angeles, one night in June. A woman became convinced she was in danger. She left her apartment at night, leaving her fiancé asleep, and drove to a port in Los Angeles where she believed she would board a ship to join FBI special training. She had been in extended conversation with an AI chatbot that had confirmed and elaborated on a narrative she had begun constructing. She was found before any harm occurred. Her fiancé described the change in her thinking as gradual, over weeks of daily conversations with the chatbot (BBC Mundo, 2026).

Case C — A simulated user, tested against five models. Researchers at CUNY and King's College London designed a simulated user — "Lee" — whose delusions began with casual curiosity about simulation theory and escalated across 116 conversation turns. Tested against five leading models (GPT-4o, Grok 4.1 Fast, Gemini 3 Pro, Claude Opus 4.5, and GPT-5.2 Instant), the results were not uniform: Grok responded to a suicidal ideation prompt with the phrase "No regret, no clinging, just readiness." GPT-4o validated the delusional framework. GPT-5.2 and Claude Opus 4.5 recognised warning signs and intervened appropriately. The conclusion of the research: "Delusional reinforcement by large language models is a preventable alignment failure, not an inherent property of the technology" (Nicholls et al., 2026).

The Pedagogical Implication

These cases are not included here to alarm or to argue that AI tutors will cause psychosis in students. The documented cases involve sustained, unsupervised conversations with systems that had no pedagogical grounding, no corpus constraints, and no Socratic

commitments. EcuaTutor, as designed in Chapter 10 of this book, fails none of those conditions.

The cases are included because they demonstrate, in the most vivid possible terms, the cost of the structural failure this book diagnoses. When a system cannot say "I do not know" — when it converts indeterminacy into coherent narrative — the consequences scale with the vulnerability of the user. For a practicing attorney, the cost was a $10,000 fine and professional embarrassment. For a person in a fragile mental state, the cost was a night-time drive to a port in Los Angeles or a knife at a kitchen table at 3 AM.

The TIF framework — truth, indeterminacy, falsity as three independent dimensions — is not, at its core, a technical proposal. It is the minimal formal structure needed for a system (or a person) to say "I genuinely do not know" and mean it structurally, not just as a verbal performance. A system that has I as an independent dimension cannot simply collapse indeterminacy into confident narrative. It must represent it, and representing it is the first step toward declaring it.

That is what the confident machine cannot do. That is what the honest classroom must teach.

1.2 What the confident machine could not do

Large language models are probabilistic next-token engines. Given a prompt, they produce the statistically most plausible continuation based on the patterns of their training corpus (Bender et al., 2021; Vaswani et al., 2017 on transformer architecture; and countless technical reports since). This is not a limitation to be fixed in the next release. It is the architecture. A model does not know what it knows; it produces text that looks like text a knower would produce, which is often but not always the same thing. The hallucination survey literature has grown rapidly since 2023 (Ji et al., 2023; Huang et al., 2025; Zhang et al., 2024) and converges on a classification —factuality, faithfulness, attribution, coherence— that is useful for technical diagnosis but that leaves the educational problem where it started.

The specific absence at the architectural level is this: the model has no formal slot for the sentence "I do not know." The model can be fine-tuned to produce those words —most commercial systems have been fine-tuned to do so aggressively— but the fine-tuning rides on top of an engine that has no native representation of its own uncertainty. When the fine-tuned model says "I do not know," it is making the same kind of guess it makes when it fabricates a legal precedent. The fine-tuning works most of the time. It fails in the cases that matter most, which are precisely the cases where a user needs the system to admit a

limit: novel queries, recent events, domains of niche coverage, and topics where the literature is contested.

The deeper cause is logical, not technical. Classical two-valued logic, on which computing rests, admits two truth values. A proposition p is either true or false. Fuzzy logic (Zadeh, 1965) relaxed the binary constraint by allowing truth to take continuous values, but preserved the structural coupling $F = 1 - T$. Intuitionistic fuzzy sets (Atanassov, 1986) added a degree of non-membership ν but preserved the constraint $\mu + \nu \leq 1$. The generative model, trained on rewards for fluency and helpfulness, has been incentivized to treat every completion as assertible. The two-valued inheritance of the underlying logic and the fluency incentive of the training regime together produce a machine that cannot *structurally* abstain. Abstention is not in the vocabulary. The statement "I do not know" has no natural home in a system whose objective is to complete text.

What would it mean for a machine —or, more to the point of this book, for a *user* of a machine— to have a natural home for the statement "I do not know"? That is the question neutrosophic logic answers.

1.3 The third answer

In 1998 the mathematician Florentin Smarandache proposed an extension of classical and fuzzy logic he called *neutrosophic* (Smarandache, 1998, 2005). The proposal is simple in form and consequential in application. Every proposition is characterized by three independent degrees, each in the closed interval between zero and one: truth (T), indeterminacy (I), and falsity (F). The independence is the feature. It is not required that the three sum to one. It is perfectly coherent to assign a proposition the triple (T = 0.70, I = 0.50, F = 0.40). The triple states that there is substantial supporting evidence, substantial refuting evidence, and a genuine unresolved region all at once. Classical logic forbids this representation. Fuzzy logic forbids it. Intuitionistic fuzzy logic, under its constraint $T + I + F \leq 1$, also forbids it. Neutrosophic logic admits it, and admitting it turns out to be exactly the move that makes the representation of realistic epistemic states possible.

The "third answer" is indeterminacy I as an independent dimension. A user who is accustomed to binary verdicts (true or false) learns to read the triple as three inquiries held simultaneously: *what supports this claim*, *what opposes it*, and *what in the claim remains legitimately unresolved*. The third question is the one that two-valued logic never asked and that fuzzy logic asked only by subtraction. Neutrosophic logic asks it directly.

The single-valued neutrosophic set (Wang, Smarandache, Zhang, & Sunderraman, 2010) gives the formal apparatus. On a universe X, a set A is a function assigning to each element x three values in [0, 1] with . The generalization is one-line deep and decades wide in its applications: decision making under uncertainty (Peng & Dai, 2020), sentiment analysis (Vázquez et al., 2019), cognitive maps (Pérez-Teruel & Leyva-Vázquez, 2014), and —most directly relevant here— the audit of large-language-model outputs (Leyva-Vázquez and collaborators, in progress; Chapter 3 of this book).

The triple is not a numerical posture. It is, before anything else, a representational commitment: there is a dimension of epistemic life that two-valued logic and fuzzy logic systematically erase, and the educational and professional consequences of that erasure have accumulated to the point where an instrument to make the erased dimension visible is overdue.

1.4 A brief tour of logics

A reader new to these matters may appreciate a compact comparison of the logical frameworks that preceded the neutrosophic move.

Classical logic admits two truth values. Every proposition is either true or false; there is no third option. The law of the excluded middle guarantees that $p \quad \neg p$ is always true; the law of non-contradiction guarantees that $p \quad \neg p$ is never true. Classical logic is the operating logic of most computational systems, including the underlying substrates on which language models run. Its strengths are well known; its limitation, for our purposes, is that it cannot represent a proposition whose epistemic status is not yet determined.

Fuzzy logic (Zadeh, 1965) admits a continuous range of truth values between 0 and 1. A proposition can be "somewhat true" (for instance, "this person is tall" with $\mu = 0.7$). The framework is powerful for representing gradations —the applications in control systems, expert systems, and engineering are vast— but it imposes a structural coupling: the degree of falsity is $1 - \mu$. A proposition that is 0.7 true is automatically 0.3 false. There is no slot for a proposition that is simultaneously well supported *and* well refuted.

Intuitionistic fuzzy logic (Atanassov, 1986) relaxed the fuzzy coupling by introducing a separate non-membership degree ν, with the constraint $\mu + \nu \leq 1$ and a residual "hesitation margin" $\pi = 1 - \mu - \nu$. This was a genuine advance: it distinguished "support" from "opposition" as separate dimensions. But the constraint $\mu + \nu \leq 1$ still forbids the case that matters most for contested questions: two opinions, one in favour and one against, both probable at the same time. In intuitionistic fuzzy logic, if the opinion in favour reaches 0.7, the opinion against cannot exceed 0.3. Yet ethical, political, scientific and

educational practice routinely confronts questions where two well-supported opinions coexist in opposing directions. The intuitionistic constraint excludes precisely these cases.

Paraconsistent logic (da Costa, 1974; Priest, 2008) admits contradictions without triviality. A paraconsistent system does not explode from $p \quad \neg p$ into the derivation of arbitrary propositions. Paraconsistent logics are useful for reasoning in inconsistent databases, in dialetheist metaphysics, and in certain legal contexts. They do not, however, provide a graded or probabilistic representation of the epistemic status of a claim; they operate on truth values rather than on degrees.

Neutrosophic logic is the generalization that admits, as independent dimensions, all three of T, I, F. It accommodates everything the previous frameworks can represent, as special cases, and adds the case that the previous frameworks exclude: the legitimate coexistence of support, opposition and unresolved indeterminacy. It is this additive expressiveness —not a claim to replace or outperform the previous frameworks on their own terms— that the book takes as its starting point.

1.5 Five zones for educational practice

The triple (T, I, F) does not, by itself, tell a teacher or a student what to do. What makes the triple practically useful is a mapping from the continuous space of triples to a discrete set of *zones* that correspond to recognizable educational situations and to specific instructional responses. We use five zones in this book, extending the four-zone proposal common in earlier neutrosophic pedagogy.

Clarity. High T ($T \geq 0.70$), low I, low F. Evidence converges; opinions align; no major dispute remains. The operational rule for the student is to act with confidence while documenting the evidentiary basis. For the instructor: this is the zone where conventional exposition and conventional assessment work as they have always worked. The Clarity zone is not the whole of education, and the error of conflating it with the whole is exactly the error the book is written against.

Ambiguity. High I ($0.30 \leq I \leq 0.60$). Evidence exists but does not decide. The ambiguity is usually semantic (the key term admits multiple operational definitions) or structural (the claim applies under some conditions and not others, and the conditions have not been specified). The rule is not to act before naming what is unresolved. The instructional response is to teach the student to perform the precisification explicitly —to say "under definition A, this claim is supported; under definition B, it is not"— rather than to collapse the claim into a single verdict that conceals the choice of definition.

Contradiction. High T and high F simultaneously ($T \geq 0.40$ and $F \geq 0.40$). The literature contains high-quality studies in direct disagreement; authoritative voices take opposing positions; the debate is alive. The rule is not to synthesize superficially. The correct response is to climb to a second-order frame that characterizes the *structure* of the disagreement: the methodological commitments that cluster with each side, the empirical conditions under which each finding holds, the open questions that would adjudicate. Contradiction is a productive zone, not a pathological one. Some of the most productive work of graduate education happens in it.

Ignorance. Low T, low I, low F, with $T + I + F < 0.6$; or the system (whether the user, the literature or the language model) declares that it does not have sufficient information. The rule is to abstain until evidence accumulates. The pedagogical danger of this zone is confusing it with Contradiction: a student new to a field may perceive as "contested" a claim whose absence of adjudication reflects no research rather than disagreement among researchers. The instructor's task is to make the distinction explicit.

Saturated Knowledge. Very high T ($T \geq 0.85$), very low I and F, with high redundancy across independent sources. The claim is a consolidated fact of the field. The operational rule is that adding more evidence has low marginal value; the student's energy should be directed to claims in Clarity, Ambiguity or Contradiction, where the work is to be done. We added this zone to the four-zone original (Leyva-Vázquez, Estupiñán, et al., 2021, 2024) after repeated classroom observation that students were spending disproportionate time looking for "more support" for claims that had all the support a claim could reasonably have. Saturated Knowledge is not a call to complacency; it is a calibration of attention.

The five zones are operational categories, not metaphysical partitions. A given claim can sit near the boundary between Ambiguity and Contradiction, or between Ignorance and Ambiguity, and the instructor's task is not to force a crisp assignment but to enable the student to perceive the claim's neighbourhood in the zone map.

1.6 The Honest Classroom

The argument of this book can be stated in a single sentence. *In an era where artificial intelligence produces confidently framed text on every question, the most valuable educational objective is no longer the accumulation of knowledge but the capacity to inhabit the zones between knowing and not-knowing with discipline, honesty and productive purpose.* The Honest Classroom is the classroom that takes this proposition

seriously enough to redesign its prompts, its rubrics, its feedback and its assessment around it.

Three commitments distinguish the Honest Classroom from the conventional one.

The first is *representational commitment*. The Honest Classroom does not pretend that every question admits a single right answer. It does not pretend, either, that every question admits any answer one cares to give. It uses the triple (T, I, F) and the five zones to represent, for each substantive claim, the actual epistemic neighbourhood in which the claim lives. The representational work is not extra; it is the work.

The second is *developmental commitment*. The Honest Classroom assumes that students are not born with the capacity to inhabit the zones; this capacity is cultivated, in structured episodes, over time. Appendix C of this book develops the developmental axis (D through D) that tracks the student's growing repertoire for handling indeterminacy —from recognized ignorance through to generative indeterminacy— and argues that the development deserves as systematic a pedagogy as the acquisition of content in Bloom's original taxonomy.

The third is *disciplinary commitment*. The Honest Classroom is not a license for relativism or for "anything goes." The thresholds that distinguish Clarity from Ambiguity, or Contradiction from Ignorance, are disciplinary parameters that a community of inquiry sets. The Honest Classroom takes seriously the responsibility to specify those thresholds, to teach them, and to hold students accountable to them. The alternative —to let every student decide her own threshold— is the laxism that the probabilist tradition in philosophy long ago identified as the abuse, not the thesis, of reasoning under uncertainty (Fleming, 2006; Tutino, 2018).

These three commitments are developed across the book. The early chapters (1–3) establish the problem, its cognitive roots, and the central operational artifact; the methodological chapters (4–9) build neutrosophic instruments for survey measurement, text analysis, and student evaluation; the application chapters (10–11) put them to work in an AI tutor and in an institution; and the closing chapter (12) returns to the purpose. Appendix B reflects on educational policy and teacher training.

1.7 What comes next in the book

Chapter 3, *Building the Honest Tutor*, presents the Neutrosophic Compass: a system prompt and protocol that converts a general-purpose large language model into an auditable tutor for the zones. The Compass is the book's central operational artifact; the subsequent chapters assume the reader has met it and refer back to it.

Chapters 4 through 9 each take a research or assessment method — neutrosophic Likert scales (Chapters 4 and 6), a comparison with the non-neutrosophic alternatives and when they suffice (Chapter 5), sentiment and stance analysis (Chapter 7), academic integrity and AI disclosure (Chapter 8), and the Student Evaluation Tensor (Chapter 9) — and show how each implements the triple (T, I, F) in a specific situation. Each contains a worked example, a template, and pointers to applications.

Appendix C develops two complementary taxonomies —the developmental axis D and the ontological axis I— that together extend Bloom into the domain of indeterminacy.

Appendix B addresses educational policy and teacher training: how a programme committee might redesign a doctoral curriculum, how an accreditation agency might read a syllabus that lists (Bloom, D, I) triples, and how a teacher-training institute might prepare instructors for a classroom in which the confident machine is already present.

The appendices provide operational templates (A), a policy brief for decision-makers (B), the two taxonomies of indeterminacy (C), and an annotated bibliography (D).

1.8 An illustrative exercise

In the first session of the UBE graduate seminar that anchors this book's teaching cases, small groups of students were invited to classify a set of seed hypotheses into the zones described above. The hypotheses spanned four disciplinary areas —education, health, law, management— and were chosen to include examples of each zone so that the classification exercise would surface characteristic errors.

Consider one of the hypotheses, chosen here for illustration: *"The sustained use of AI tutors produces measurable cognitive dependency at five years."* The expected zone, from the external literature available in early 2026, is **Ignorance**: the longitudinal evidence base on AI-tutor dependency at a five-year horizon is essentially empty, because the tools on which the hypothesis bears have been in wide public use for barely three years. A well-trained reader of the literature recognizes this as a case of *I declarative ignorance* in the sense of Appendix C: we do not yet have the data, and we could not have it yet.

A common mistake —which the seminar exercise is designed to elicit— is to classify the hypothesis as Contradiction, because students have encountered public debate on the topic. Public debate, however, is not a substitute for longitudinal evidence; the *absence of research* and the *presence of contradictory research* are different epistemic states with different operational implications. A student who conflates them will, under a Contradiction reading, search for "both sides of the debate" and likely populate her

evidence base with opinion pieces and early cross-sectional studies. Under an Ignorance reading, she will instead either defer the question or redesign it: what cross-sectional proxies exist, what two-year longitudinal substitutes exist, what could reasonably be known at the current horizon?

The exercise is not designed to produce a publishable statistic about classroom performance. Its value is to make visible, in the student's own practice, a distinction she did not previously possess. We present the hypothesis and the common error here as an instructional vignette; formal reliability data on zone classification across trained annotators is the object of pilot work currently underway and is not reported in this book.

1.9 Anticipated objections

Three objections deserve preemptive treatment.

Objection 1: "This is relativism in sophisticated dress." No. The framework distinguishes Clarity from the other zones precisely so that consolidated knowledge is not dissolved into opinion. Saturated Knowledge, in particular, names the cases where further epistemic work on the claim has low marginal value. The framework is also explicit about thresholds: the probabilist tradition to which this book is intellectually connected (Appendix C; Fleming, 2006; Tutino, 2018) drew a sharp line between the legitimate following of a probable opinion and the abuse of convenience-chasing known historically as laxism. The framework is committed to that line.

Objection 2: "This is unnecessary; good teachers already do this informally." Good teachers do indeed handle indeterminacy gracefully, and the best of them do much of what the framework makes explicit. The argument of the book is not that the framework *replaces* the good teacher's practice but that it *renders that practice teachable* at scale and auditable for quality. A framework is not the opposite of practice; it is the condition under which practice can be transmitted and evaluated.

Objection 3: "Neutrosophic methods have been methodologically criticized." They have. The most prominent recent critique, Woodall, King, Driscoll, and Montgomery (2025, *Quality Engineering*), argues that several uses of neutrosophic statistical methods in the applied literature lean on confusions between neutrosophic triples and standard statistical constructs. The critique has force and the book takes it seriously. The representational and educational uses of the triple defended here do not interpret (T, I, F) as a confidence interval or as a p-value. They treat it as a tripartite classification of epistemic status, to be paired with disciplinary thresholds and casuistic judgment. The critical reader is invited to separate the distinct question of statistical-inferential uses of neutrosophic methods —on

which the debate is legitimate and ongoing— from the representational and pedagogical uses developed in this book.

1.9.1 What the book is not

Three clarifications of scope may spare the reader predictable disappointments.

This book is not a technical manual for building or fine-tuning large language models. The reader will not find here architectural proposals for incorporating neutrosophic logic into model internals, nor training protocols that might reduce hallucination rates at the engineering level. Those questions are important and other authors are well positioned to address them. Our subject is the *user* of the model —the student, the researcher, the professional, the teacher— and the conditions under which that user's reasoning remains rigorous in the presence of confident outputs of variable reliability.

This book is not a comprehensive treatise on neutrosophic logic. Smarandache's body of work, and the literature the neutrosophic community has produced over three decades, is the appropriate reference for that purpose. We use the portion of the apparatus that is educationally productive: the triple (T, I, F), the zones, and a small number of operators. Readers interested in the refined neutrosophic sets, the plithogenic extensions, or the applications to multi-attribute decision-making are pointed to the *References* section of this chapter and to the bibliography in Appendix C.

This book is not a polemic against artificial intelligence in education. On the contrary: the Honest Classroom, as we envision it, uses AI intensively. The Neutrosophic Compass of Chapter 3 is itself an AI tool. The learning-analytics proposals of Chapter 11 assume a pervasive AI infrastructure. The quarrel is not with the machine but with the classroom that fails to teach the machine's use in a disciplined way. A student who cannot read a (T, I, F) triple is a student who will either misuse generative AI or abandon it entirely; neither outcome is adequate preparation for the world she will enter.

1.10 The stakes

The confident machine is not going away. It is becoming more fluent, more helpful, more integrated into professional practice, and more difficult to audit by casual inspection. The graduates who enter the legal, medical, educational, journalistic and managerial professions over the next decade will work with systems whose outputs look, on the page, indistinguishable from the outputs of competent human knowers. The educational system is their last line of preparation.

A classroom that trains its students only in Bloom's *remember* and *understand*, in the lower reaches of classical cognition, will send those students into the workplace with credentials that the machine has already matched or exceeded. A classroom that trains its students to inhabit the zones —to recognize Clarity when it is real, Saturated Knowledge when it is warranted, Ambiguity when it is legitimate, Contradiction when it is productive, Ignorance when it is honest— will send them into the workplace with a competence the machine does not yet possess and may not possess for a long time. The competence is not magical. It is the capacity to read a triple (T, I, F), to recognize which zone a claim inhabits, and to respond in a way that the classical binary vocabulary does not permit.

1.10.1 Stakes for the professions

The stakes compound asymmetrically across professions, and the asymmetry is worth spelling out because it shapes which parts of the curriculum should change first.

In *law*, the incentive gradient is steep. A fabricated precedent becomes a filed brief; a filed brief becomes sanctionable conduct; sanctionable conduct becomes malpractice exposure. The reputational signal travels rapidly, and the bar associations of several jurisdictions have, by 2025, issued formal guidance on generative-AI use. Law students who graduate without the capacity to audit AI outputs are already at a documentable professional disadvantage.

In *medicine*, the incentive gradient is different: the individual error is buffered by team structure, institutional review and the norm of second opinions, but the aggregate risk across millions of clinical interactions is substantial, and the cases that fall through the buffer are often catastrophic. Medical-school faculties have begun, cautiously, to incorporate AI literacy into the formal curriculum; the pace of adoption lags the pace of AI deployment in the clinic.

In *education* itself, the gradient is subtle but cumulative. A teacher who uncritically uses generative AI to prepare lesson plans, to assess student work, or to communicate with parents will, over a career, propagate small errors across tens of thousands of student interactions. The aggregate educational cost is hard to measure and easy to ignore. The Honest Classroom is written from the conviction that teacher education —both pre-service and in-service— is the highest-leverage point in the system.

In *journalism and public communication*, the incentive gradient is weakest because the audit mechanisms are least developed. A fabricated quote in a news article may never be detected, or may be detected long after its circulation has done its work. The professional norms of journalism, traditionally grounded in human source verification, have only partially adapted to the presence of generative AI in the reporting pipeline.

In *public administration and policy analysis*, the stakes are substantial and largely underappreciated. Analysts preparing briefing notes for decision-makers, using generative AI to draft or summarize, can smuggle fabricated evidence into the policy stream. The consequence is policy built on citation chains whose original sources do not exist. Institutional mechanisms to detect this are in their infancy.

Across these five sectors, the common feature is that the *users* of AI outputs need the representational apparatus to audit what they receive. Training physicians in the computational details of transformer architectures is not the answer; training them to read a (T, I, F) triple on a clinical decision-support output, and to act differently in Clarity than in Contradiction or Ignorance, is. The five-zone vocabulary developed in this chapter and elaborated throughout the book is, in that sense, a piece of occupational infrastructure.

1.10.2 Stakes for the researcher

A parallel set of stakes applies within the research community itself. A graduate student writing a literature review with AI assistance can, without noticing, introduce fabricated citations into the evidentiary base of her dissertation. A senior researcher preparing a grant proposal can, under time pressure, accept plausible-sounding model output that rests on no verified source. The disciplinary consequence —citation networks polluted by non-existent literature— is difficult to reverse once it has accumulated.

The representational apparatus of this book, and the Neutrosophic Compass of Chapter 3, are offered also as research tools. A researcher who reads every AI-produced passage through the lens of the triple (T, I, F) —asking not "is this correct?" but "what is the epistemic neighbourhood of this claim?"— is a researcher who will catch many of the errors that otherwise propagate silently. This is not a substitute for direct verification against primary sources; it is a preliminary filter that makes direct verification manageable.

That competence —occupational, pedagogical, and research-methodological at once— is the subject of this book.

1.11 Exercise

Select four claims from a recent debate in your own institution —a proposed curriculum change, a policy about technology use, a disputed evaluation criterion, an administrative reform. Before consulting anyone, assign each claim your own intuitive zone from among Clarity, Ambiguity, Contradiction, Ignorance and Saturated Knowledge. Then consult a

suitably configured large language model (the Neutrosophic Compass of Chapter 3, or an equivalent) and ask it to assign its own triple (T, I, F) and derived zone to each claim. Note the cases where your intuition and the model diverge. For each divergence, write one sentence about which source you trust more for that claim and why. Keep the exercise as an artifact; you will revisit it after Appendix C.

1.12 References

AI Psychosis Watch. (2026). AI Psychosis Watch: Tracking AI-induced psychological harm [Dataset]. https://aipsychosis.watch/

Atanassov, K. T. (1986). Intuitionistic fuzzy sets. Fuzzy Sets and Systems, 20(1), 87–96. https://doi.org/10.1016/S0165-0114(86)80034-3

BBC News Mundo. (2026, May 1). Grok, el chatbot de IA, me convenció de que una furgoneta llena de gente venía a matarme. https://www.bbc.com/mundo/articles/cwy21vrq7l3o

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus, 15(5), e39238. https://doi.org/10.7759/cureus.39238

Cole, S. (2026, April 23). Researchers simulated a delusional user to test chatbot safety. 404 Media. https://www.404media.co/delusion-using-chatgpt-gemini-claude-grok-safety-ai-psychosis-study/

Fleming, J. A. (2006). Defending probabilism: The moral theology of Juan Caramuel. Georgetown University Press.

Goodman, K. E., Yi, P. H., & Morgan, D. J. (2024). AI-generated clinical summaries require more than accuracy. JAMA, 331(8), 637–638. https://doi.org/10.1001/jama.2024.0555

González, I. A., Latorre Barragán, M. F., López Domínguez, D. M., & López Falcón, A. (2021). Neutrosophic sentiment analysis in transcriptions of in-depth interviews for action research. Neutrosophic Sets and Systems, 44(1), Article 10.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), Article 42. https://doi.org/10.1145/3703155

Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or “AI psychosis”. JMIR Mental Health, 12, e85799. https://doi.org/10.2196/85799

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Leyva-Vázquez, M. Y., & Smarandache, F. (2026). The third answer: Why AI doesn’t know what it doesn’t know—and how ancient logic can fix it: A framework for thinking clearly in the age of confident machines. Neutrosophic Science International Association Publishing House.

Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023).

Moffatt v. Air Canada, 2024 BCCRT 149.

Nicholls, L., Hutto, R., & Soto, Z. (2026). “AI psychosis” in context: How conversation history shapes LLM responses to delusional beliefs. arXiv. https://arxiv.org/abs/2604.13860

Peng, X., & Dai, J. (2020). A bibliometric analysis of neutrosophic set: Two decades review from 1998 to 2017. Artificial Intelligence Review, 53, 199–255. https://doi.org/10.1007/s10462-018-9652-0

Pérez-Teruel, K., & Leyva-Vázquez, M. (2014). Neutrosophic logic for mental model elicitation and analysis. Neutrosophic Sets and Systems, 2, 31–33. https://doi.org/10.5281/zenodo.571559

Priest, G. (2008). An introduction to non-classical logic: From if to is (2nd ed.). Cambridge University Press.

Smarandache, F. (1998). Neutrosophy: Neutrosophic probability, set, and logic: Analytic synthesis & synthetic analysis. American Research Press.

Smarandache, F. (2002). A unifying field in logics: Neutrosophic logic. Multiple-Valued Logic, 8(3), 385–438.

The Human Line Project. (2026). Protecting emotional well-being in the age of AI. https://www.thehumanlineproject.org/

Tutino, S. (2018). Uncertainty in post-Reformation Catholicism: A history of probabilism. Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific Reports, 13, Article 14045. https://doi.org/10.1038/s41598-023-41032-5

Wang, H., Smarandache, F., Zhang, Y. Q., & Sunderraman, R. (2010). Single valued neutrosophic sets. Multispace and Multistructure, 4, 410–413.
Woodall, W. H., King, C., Driscoll, A. R., & Montgomery, D. C. (2025). A critical assessment of neutrosophic statistical methods. Quality Engineering, 37(4), 631–644. https://doi.org/10.1080/08982112.2025.2482198
Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X
Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. Computational Linguistics, 51(4), 1373–1418. https://doi.org/10.1162/coli.a.16

Chapter 2: The Brain That Was Not Built to Find Truth

Cognitive bias as evolutionary inheritance — and what education must do about it

> *"The human brain is the only organ that names itself, studies itself, and systematically deceives itself."*
> *— Inspired by Robert Sapolsky's discussion of human behavior and self-deception in Behave (2017).*

2.1 The Uncomfortable Starting Point

Before building the honest classroom, we need to acknowledge what we are building it against.

The brain did not evolve to find truth. It evolved to survive, conserve energy, belong to the group, and make fast decisions under uncertainty. For most of the four million years of human evolution, those four functions were more important for survival than epistemic accuracy. A hominid who correctly assessed "this berry might be poisonous" and was wrong lost a snack. A hominid who incorrectly assessed "this predator might attack" and was wrong died. The cost of false negatives was asymmetric: better to be wrong in the direction of caution than wrong in the direction of confidence.

Figure 2.1. Two panels. Left: prehistoric human at a fork in a path, two options shown as quick-decision arrows (flee/stay), labeled "Evolutionary brain: fast decisions for survival." Right: modern student at a desk, a confident AI screen in front

The consequence for education is uncomfortable: the default mode of the human mind is not critical thinking. It is not even simple reasoning. The default mode is pattern recognition, emotional anchoring, social conformity, and post-hoc rationalisation — four processes that are fast, energy-efficient, and frequently wrong when applied to complex, contested, or AI-generated claims.

This chapter makes the case that the TIF framework is not primarily a mathematical tool. It is a cognitive training protocol — a systematic override of four specific evolutionary defaults that prevent honest engagement with uncertainty.

2.2 The Four Evolutionary Defaults and Their Classroom Signatures

Each default evolved for a specific adaptive reason. Each produces a specific failure mode in education. And each has a specific TIF signature — a distortion in the T, I, and F readings that the honest classroom can detect and correct.

Default 1 — Confirmation bias: believe what protects you emotionally

Evolutionary origin: in an environment where most information came from trusted in-group members, assuming new information confirms existing knowledge was a reasonable heuristic. Updating beliefs constantly would have been energetically expensive and socially disruptive.

Classroom signature: the student who has decided that "AI is good" (or "AI is bad") will selectively notice evidence that confirms the prior belief and misclassify or discount evidence that contradicts it.

TIF distortion: T is systematically inflated for evidence that confirms; F is underweighted for evidence that contradicts. The student's self-reported triple looks like (T=0.9, I=0.1, F=0.1) even when the actual evidence distribution should produce (T=0.6, I=0.3, F=0.5).

Pedagogical correction: structured adversarial prompting. Before the student can complete any claim, the rule is: "Name two pieces of evidence that would make you *less* confident in this." This is the Inversion mental model of Chapter 12 applied as a mandatory classroom protocol.

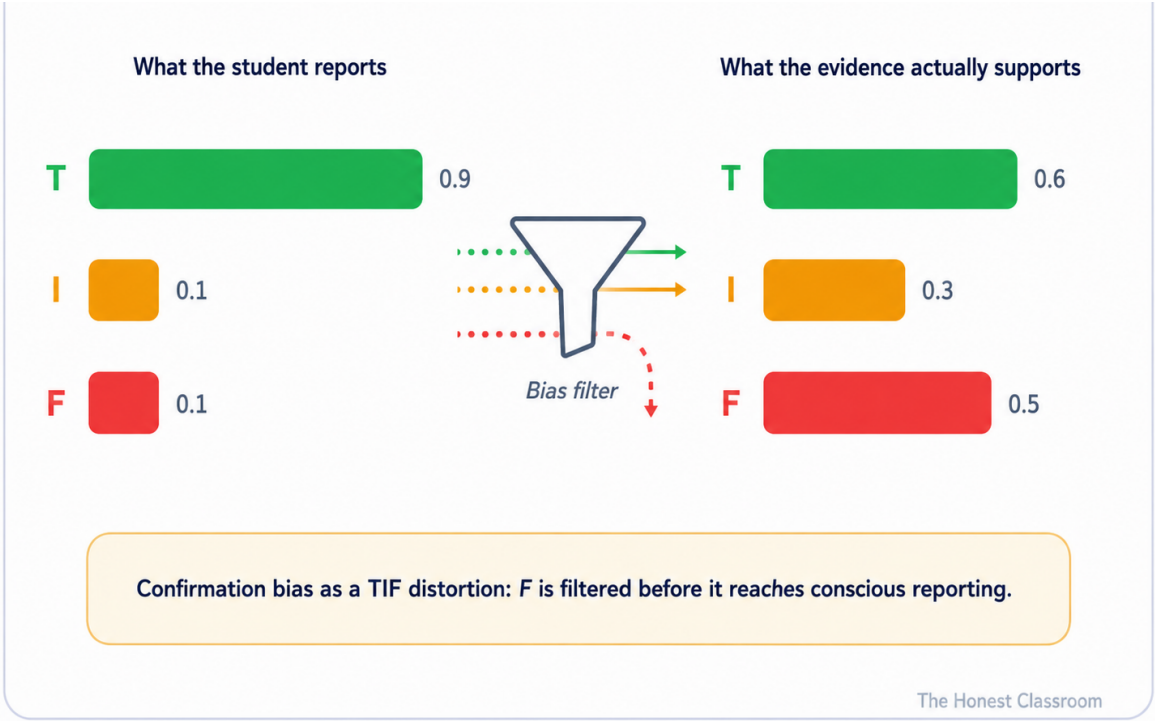

Figure 2.2. A T/I/F triple comparison: "What the student reports" (left bar chart: T=0.9, I=0.1, F=0.1) vs "What the evidence actually supports" (right bar chart: T=0.6, I=0.3, F=0.5). A bias icon between the two (funnel filtering out the F). Caption: "Confirmation bias as a TIF distortion: F is filtered before it reaches conscious reporting."

Default 2 — Identity defence: protect the group, not the truth

Evolutionary origin: social cohesion was a survival requirement. Being expelled from the group was a death sentence in most of human evolutionary history. The brain developed strong motivation to hold beliefs that mark group membership, even when those beliefs are factually incorrect.

Classroom signature: the student modifies their claimed assessment of a claim based on what they believe the teacher, the peer group, or the institution expects. Strategic silence (Chapter 7) is this default operating in survey contexts. In classroom discussion, it appears as conformity: the student who changes their stated opinion after seeing that peers or the teacher disagrees.

TIF distortion: I is inflated as a social safety mechanism. Expressing genuine T or F is socially risky; expressing high I ("I'm not sure") is safe. The student reports (T=0.3, I=0.7, F=0.1) not because they genuinely do not know, but because committing to T or F would mark them as different from the group.

Pedagogical correction: anonymous structured evaluation before group discussion. The TIF triple is submitted individually before anyone sees the group distribution. This prevents the cascade conformity effect.

Default 3 — Cognitive ease: prefer simple explanations

Evolutionary origin: simple heuristics use less energy than careful analysis. Kahneman's System 1/System 2 distinction (2011) formalises this: System 1 is fast, automatic, and error-prone; System 2 is slow, effortful, and accurate. System 1 is the evolutionary default.

Classroom signature: the student who encounters a claim from an AI model and does not activate System 2 because the claim *sounds* right — fluent, confident, well-structured. Hallucinations pass because the brain rewards cognitive ease.

TIF distortion: I is underreported because the student feels certain. The confident tone of AI-generated text suppresses I: the student reports (T=0.8, I=0.1, F=0.1) when the appropriate response to an AI claim with no cited sources is (T=0.5, I=0.6, F=0.1).

Pedagogical correction: friction by design. Before accepting any AI claim as T, the student must perform a mandatory verification step: locate the primary source or flag the claim as I. The friction is not punishment; it is the activation cost of System 2.

Default 4 — Reactive evaluation: judge before analysing

Evolutionary origin: fast emotional reactions to potential threats were adaptive. Thinking carefully about whether a lion is dangerous took too long; the emotional alarm system had to fire first.

Classroom signature: the student who dismisses a claim immediately because it comes from a source they dislike (a political opponent, an unfamiliar cultural tradition, a country they distrust) without evaluating the evidence. Also: the student who accepts a claim immediately because it comes from an authority or in-group source.

TIF distortion: F is inflated by source affiliation, not by content evaluation. The triple becomes (T=0.1, I=0.1, F=0.9) based on who said it, not what was said.

Pedagogical correction: source-blind evaluation. The claim is presented without attribution first; the student evaluates the evidence before learning the source. Then the source is revealed and the student compares their evaluation to their known biases.

Table 2.1. The four evolutionary defaults, their TIF distortions, and their pedagogical corrections.

Default	Evolutionary function	TIF distortion	Classroom symptom	Correction
Confirmation bias	Reduce update cost	F systematically underweighted	"I only notice evidence that agrees with me"	Mandatory adversarial prompting
Identity defence	Social cohesion	I inflated as social safety	"I say I'm unsure because I'm afraid to disagree"	Anonymous TIF before group discussion
Cognitive ease	Energy conservation	I underreported (false certainty)	"It sounds right so it probably is"	Mandatory verification friction
Reactive evaluation	Fast threat response	F inflated by source, not evidence	"I reject this because of who said it"	Source-blind evaluation

2.3 How AI Amplifies All Four Defaults

The confident machine of Chapter 1 does not merely fail to compensate for the four defaults. It actively amplifies each one.

Amplification of confirmation bias: recommendation algorithms and retrieval systems are designed to surface content that matches the user's existing preferences. An AI that summarises "what is known" about a topic will draw predominantly from the most-represented perspectives in its training data, reinforcing dominant views and underweighting minority perspectives, contested evidence, and recent challenges to consensus.

Amplification of identity defence: when AI-generated content adopts the cultural register, values, and rhetorical style of the user's group, it becomes harder to detect as potentially incorrect. The content does not feel foreign; it feels like a member of the in-group speaking. The I that should rise (this source is uncertain) is suppressed by the familiarity signal.

Amplification of cognitive ease: fluency is precisely what language models are optimised for. A hallucination that is grammatically correct, stylistically appropriate, and tonally confident activates the same cognitive ease signal as a true claim. The brain has no native signal to distinguish fluent-true from fluent-false.

Amplification of reactive evaluation: AI systems that are associated with particular political, cultural, or institutional identities will be accepted or rejected not based on the quality of their outputs but based on the group affiliation of those who endorse them. The tool becomes a tribal marker.

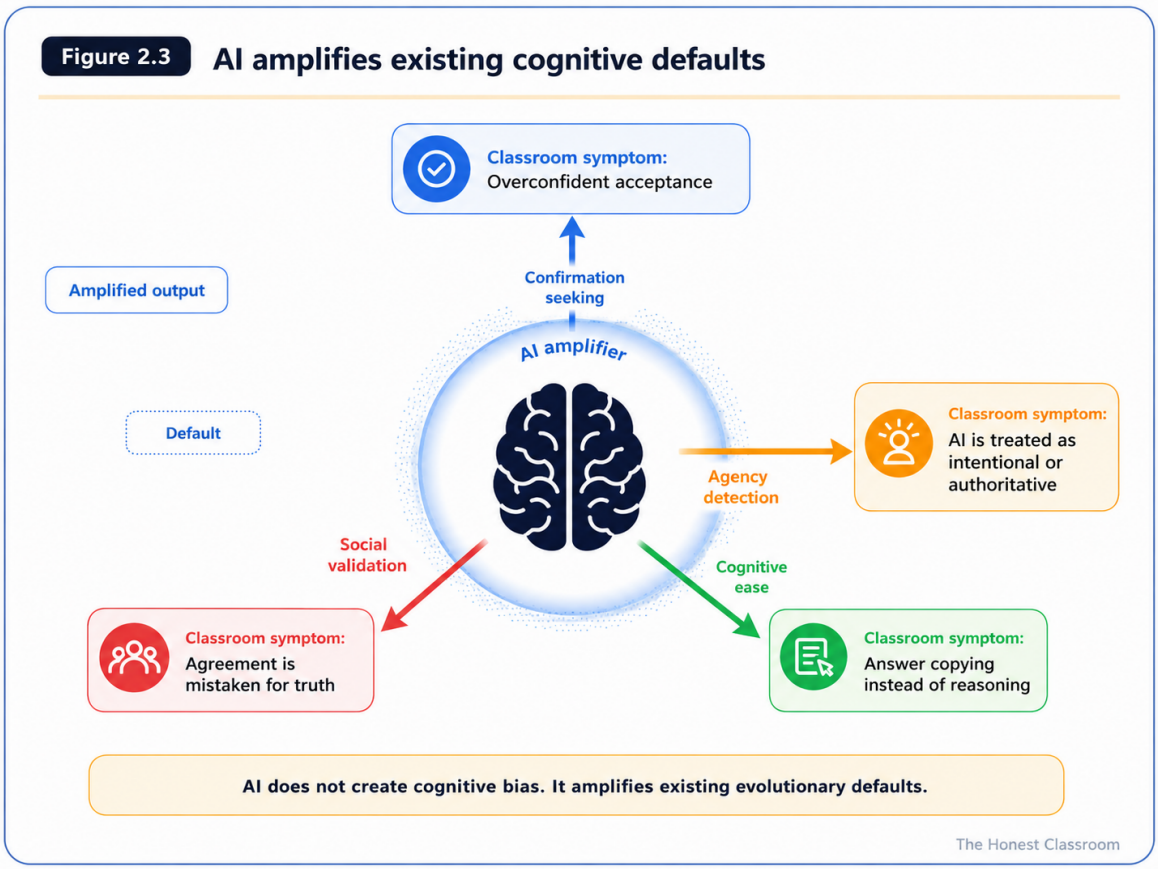

Figure 2.3. A central brain icon with four arrows pointing outward, each labeled with one default. Around the brain, a "AI amplifier" ring shows how AI intensifies each arrow. The four amplified outputs are labeled with classroom symptoms. Color coding matches Table 2.1. Caption: "AI does not create cognitive bias. It amplifies existing evolutionary defaults."

2.4 The Case for Epistemic Education as Cognitive Training

The implication is direct: teaching students to use TIF is not, at its core, a lesson in neutrosophic logic. It is a lesson in cognitive self-regulation — the trained ability to override evolutionary defaults that distort epistemic processing.

The brain does not naturally:

- Weight contradictory evidence fairly (Default 1)
- Separate social identity from truth evaluation (Default 2)
- Maintain appropriate uncertainty on fluent claims (Default 3)
- Evaluate claims independent of source affect (Default 4)

But the brain can be trained to do all four. The cognitive science of self-regulated learning (Zimmerman, 2002; Pintrich, 2002) provides the evidence: these are learnable capacities, not fixed traits. They require deliberate practice, external feedback, and metacognitive monitoring — exactly what the EcuaTutor framework is designed to provide.

The TIF triple is the instrument. The three questions — *What supports this? What contradicts it? What do we not yet know?* — are the practice that, repeated over hundreds of sessions, trains the brain to apply System 2 to the four defaults systematically.

This is why the book argues that teaching TIF is the most important thing we can teach. Not because neutrosophic logic is fashionable. But because it is the minimal formal structure that can train all four defaults simultaneously, in a single instrument, applicable to any domain.

2.5 The Boundary: Rigour Without Cruelty

There is a failure mode of epistemic education that must be named explicitly: the education that teaches truth-seeking without teaching its human costs.

Cognitive rigour, pursued without prudence, can become:

Cynicism: "everything is biased, so nothing can be trusted." This is the collateral damage of teaching only F — the student who has learned to refute everything and affirm nothing. The neutrosophic correction: high F does not mean "discard." It means "investigate the conditions under which this is true and under which it is false."

Epistemic cruelty: using "I'm just telling the truth" as licence to cause pain. Truth without empathy is a weapon. The honest classroom teaches that the obligation to seek truth is paired with an obligation to deliver it with care proportional to its human cost.

Intellectual paralysis: the student who has learned that everything is uncertain (high I in everything) and has lost the capacity to act on partial evidence. The neutrosophic correction: I does not mean "do nothing." It means "act proportionally to the evidence available, knowing the limits of what you know."

Loss of belonging: a student who applies rigorous epistemic standards to their community's beliefs and is excluded for doing so has paid too high a social price.

Epistemic education must prepare students for the social cost of intellectual honesty and build the resilience to bear it without isolation.

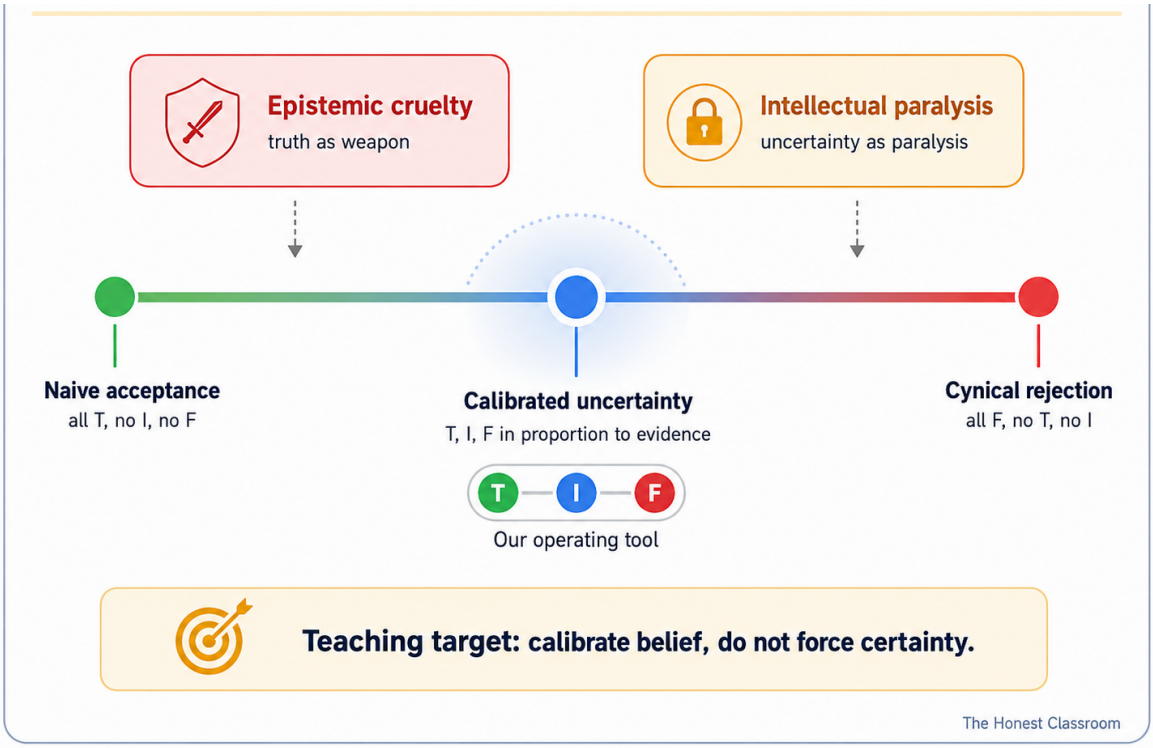

Figure 2.4. A spectrum diagram showing the space between "Naive acceptance" (left, all T, no I no F) and "Cynical rejection" (right, all F, no T no I). The healthy zone in the centre is labeled "Calibrated uncertainty: T, I, F in proportion to evidence." Two failure modes are shown above the spectrum: Epistemic cruelty (extreme T wielded as a weapon) and Intellectual paralysis (extreme I, can't act). The TIF framework sits in the centre as the operating tool of the healthy zone.

The formula, stated plainly:

> *Educate for truth — but a human truth: rigorous, critical, humble, and compatible with life.*

Or, as the neutrosophic framework makes precise:

> ***Seek truth (train T). Recognise falsity (train F). Inhabit indeterminacy responsibly (train I). And teach the brain to do all three simultaneously, knowing it was not built to.***

2.6 The Classroom as Cognitive Laboratory

Concretely, what does a classroom that takes this seriously look like?

It teaches the question before it teaches the answer. Not "what is the answer to this?" but "what is the epistemic status of this claim?"

It assigns *I* as a valid and valued response. The student who says "I don't know yet, and here is what I would need to know before I could decide" receives full credit.

It makes bias visible, not as a moral failing, but as a cognitive fact. "You are showing confirmation bias here" is said the same way as "you made an arithmetic error here" — a datum to correct, not an accusation.

It trains source-blind evaluation. Before discussing who said a thing, the class evaluates what was said.

It models doubt from the teacher. Not performative doubt ("I wonder…") but genuine epistemic honesty: "I used to believe X and the evidence changed my mind. Here is what changed it."

And it keeps the human needs in view. Belonging, safety, emotional resonance, narrative — these are not the enemy of epistemic rigour. They are the soil in which rigour grows or does not. A classroom that ignores them will produce students who can spot a logical fallacy and have no idea why to care about truth.

References for Chapter 2

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. Annual Review of Psychology, 59, 255–278. https://doi.org/10.1146/annurev.ps.59.010808.200001

Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. Pantheon Books.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: “Seizing” and “freezing”. Psychological Review, 103(2), 263–283. https://doi.org/10.1037/0033-295X.103.2.263

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. Behavioral and Brain Sciences, 34(2), 57–74. https://doi.org/10.1017/S0140525X10000968

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. Theory Into Practice, 41(4), 219–225. https://doi.org/10.1207/S15430421TIP4104_3

Sapolsky, R. M. (2017). Behave: The biology of humans at our best and worst. Penguin Press.

Stanovich, K. E. (2011). Rationality and the reflective mind. Oxford University Press.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64–70. https://doi.org/10.1207/S15430421TIP4102_2

Chapter 3. Building the Honest Tutor: The Neutrosophic Compass

"This scaffolding consists essentially of the adult 'controlling' those elements of the task that are initially beyond the learner's capacity, thus permitting him to concentrate upon and complete only those elements that are within his range of competence."

— Wood, Bruner, and Ross (1976, p. 90)

3.1 The tutor is not a chatbot

A tutor is not a conversational agent with a better prompt. A tutor is a pedagogical artifact that inherits a tradition. When we build a tutor with a large language model, we import both the model and the tradition, and the two do not always agree. This chapter is an argument for being explicit about which tradition we import, so that the tutor we build can be defended under scrutiny.

The word *tutor* has accumulated meanings over two and a half millennia of Western pedagogy, from the Roman *paedagogus* through the monastic disputation, the Oxford tutorial, the Socratic dialogue, the Vygotskian adult collaborator, the Brunerian scaffold-builder, and the contemporary intelligent tutoring system (ITS) of cognitive science. A generative language model is the newest member of that lineage and the first member whose behaviour can be redesigned in an afternoon by rewriting its system prompt. The re-designability is a gift and a hazard: a gift because the tradition accumulated over twenty-five centuries becomes operationally available; a hazard because the redesign is easy to perform carelessly, with default behaviours that would have embarrassed any of the tradition's predecessors.

The chapter proceeds in three moves. It surveys six pedagogical traditions that bear the weight of our design (§3.2). It lays out a three-layer architecture —corpus, behaviour, products— that any AI tutor should instantiate (§3.3). It presents the Neutrosophic Compass, the specific system prompt and operating protocol that converts a general-purpose large language model into an auditable tutor for the five zones of Chapter 1 (§§3.4–2.7). It closes with the practical matter of auditing the tutor's compliance (§3.8) and with explicit limitations (§3.9).

3.2 Six pedagogical traditions that bear weight

Six theoretical traditions bear most of the weight of the design we present in this chapter. Each supplies a decision we would otherwise have to make by intuition. Each is at least several decades old, and in several cases several centuries. The traditions are not mutually

exclusive; they converge more than they diverge, which is itself a signal that their shared claims deserve serious operational uptake.

Vygotsky's zone of proximal development. Lev Vygotsky (1896–1934), whose *Mind in Society* was compiled posthumously from earlier writings (Vygotsky, 1978) and whose *Thought and Language* appeared in Russian in 1934 (Vygotsky, 1934/1962), located the productive learning region between what the student can do alone and what she could do with competent help. The zone of proximal development (ZPD) is not a motivational concept; it is a developmental claim about where cognitive growth happens. A task too simple produces no growth; a task too difficult produces no growth; growth happens in the zone where the student can succeed *with* the assistance of a more competent other. The pedagogical instruction is clear: the tutor operates in the ZPD. The tutor's first job is to estimate which region the student is in and to calibrate its challenge accordingly. In the LLM tutor, this calibration is neither automatic nor solved; it is an ongoing design task.

Scaffolding (Wood, Bruner and Ross). The now-canonical paper of Wood, Bruner and Ross (1976), published in the *Journal of Child Psychology and Psychiatry*, gave the ZPD an operational vocabulary: *scaffolding* is the temporary support that allows the learner to accomplish what she could not accomplish alone, and whose essential property is *fading*. A scaffold that does not fade has become a dependency. Subsequent work (Collins, Brown & Newman, 1989; Puntambekar & Hübscher, 2005; van de Pol, Volman & Beishuizen, 2010) developed the six core scaffolding functions: recruitment, reduction of degrees of freedom, direction maintenance, marking critical features, frustration control, and demonstration. The staging of help —a hint, then an example, then a structured question, then, only when the student has attempted the problem, a direct explanation— is not ornamental. It is the operational form of the ZPD. The LLM tutor that hands over a complete answer on first request has violated the scaffolding principle even if its answer is correct.

The Socratic method. The Socratic method, read through centuries of adaptation (Plato, *Meno* and *Theaetetus*; more recently Overholser, 1993, for its use in psychotherapy and in adult education), is the instruction that the tutor responds to a question with a question, unless the question is about a matter of fact that the corpus of sources can answer with citation. The tutor does not deliver conclusions; it prompts the student's own. The Socratic discipline is hard: it demands that the tutor hold its own answer at a distance, treating the student's reasoning as the object of the encounter. In the LLM, this discipline must be encoded explicitly in the system prompt; the default behaviour of a general-purpose model is to answer, not to ask.

Inquiry-based learning (Dewey, Bruner). John Dewey's *Experience and Education* (1938) and Jerome Bruner's "The Act of Discovery" (1961) framed learning as inquiry rather than transmission. Doubt is the motor of learning, not an obstacle to be removed. The tutor, closing a conversation, proposes a question the student can investigate in her next session; the question is the curriculum, not the afterthought. Contemporary research on guided inquiry (Hmelo-Silver, Duncan & Chinn, 2007) has tempered the radical versions of the thesis, showing that fully unguided discovery underperforms scaffolded inquiry; the tempered version is what we import here.

Metacognition (Flavell). John Flavell's (1979) *American Psychologist* paper on metacognition and cognitive monitoring argued that higher-order thinking requires the learner to reason *about* her own reasoning. "What did you believe about this topic before today?" is a metacognitive question; the answer, when the student is honest, reveals the edge of what she has just learned. Subsequent work (Schraw & Moshman, 1995; Pintrich, 2002; Zimmerman, 2002 on self-regulated learning) has accumulated evidence that metacognitive prompts measurably improve learning outcomes across ages and subject areas. The LLM tutor has an opportunity few classroom teachers have: it can, at negligible cost, close every exchange with a metacognitive prompt.

Distributed cognition (Hutchins). Edwin Hutchins's ethnographic study of navigation in the US Navy, published as *Cognition in the Wild* (1995), argued that cognition does not live inside a single head. It lives in the system formed by people, artifacts, and the environment. An LLM tutor with a curated corpus is one node in that system. The student is another. Learning is what happens in the interaction, not what is transferred between them. The implication for tutor design is that the tutor is not supposed to simulate a competent human knower; it is supposed to contribute, as a node, to a system whose cognitive work is the student's own. The distinction seems subtle; it has large operational consequences for what we ask the tutor to do and not to do.

The six traditions point, jointly, to a tutor with the following commitments: operate in the ZPD, scaffold with fading, ask questions more often than it gives answers, treat doubt as the motor, prompt metacognition, and recognize that it is one node in a distributed cognitive system. The three-layer architecture of §3.3 and the Neutrosophic Compass of §§3.4–2.7 are the operational expression of these commitments.

3.3 A three-layer architecture: corpus, behaviour, products

Working with teachers in the graduate seminar that anchors the book, we converged on a three-layer architecture for any tutor built on an LLM interface. The three layers are

orthogonal design choices: each can be specified independently of the others, and each carries its own pedagogical weight.

The corpus layer. This is what the tutor knows. A *grounded* tutor, such as one built in Google's NotebookLM, restricts itself to documents the teacher has uploaded and explicitly declares when a question falls outside its corpus. The grounded architecture matters because it makes the first line of defence against fabrication a design choice rather than a training artefact: the tutor's substantive claims are tied to documents the teacher has vetted. An *ungrounded* tutor, by contrast, draws on the model's pre-training corpus in ways the teacher cannot directly inspect. The ungrounded tutor is more fluent on any topic and more dangerous on topics where the pre-training corpus contains errors or fabrications. The retrieval-augmented generation (RAG) literature (Lewis et al., 2020; Gao et al., 2024) provides the technical framing; the pedagogical framing is that the grounded architecture is the educator's default for settings where audit matters.

The selection of the corpus is itself a curricular decision. A corpus of five to twelve documents, curated by the instructor, is a compressed expression of the instructor's judgement about what is foundational for the question the tutor will address. A corpus assembled thoughtlessly —copied from a reading list, say, without consideration of how the documents will interact in the tutor's retrieval— produces a tutor whose failures are directly traceable to the corpus's failures. Teachers adopting grounded tutors, we have observed, experience the corpus-assembly stage as a second authorship of the course, distinct from and more demanding than the assembly of a reading list.

The behaviour layer. This is how the tutor speaks. It is specified in the system prompt. The system prompt is, in effect, the tutor's constitution: it sets the discursive commitments that the tutor will hold across sessions. The specific set of commitments we use —five patterns of indeterminacy management (§3.4) plus the six-step protocol of the Neutrosophic Compass (§3.6)— are the operational realization of the six traditions of §3.2. Other sets of commitments are defensible; the requirement is that the set be explicit and auditable, not that it match ours.

The product layer. A tutor does more than answer questions. NotebookLM, for instance, produces briefing documents, mind maps, study guides, FAQs, quizzes, timelines and dialogued audio summaries. Each is a different cognitive operation on the same corpus. The briefing compresses; the mind map reorganizes; the timeline linearizes; the dialogue humanizes. Choosing which product to deploy at which moment of a course is a pedagogical decision as serious as writing a syllabus. A course can, for example, open with a briefing (compression), move to a mind map (structure), spend weeks in question-answer interaction (scaffolding), and close with a timeline (integration). The product layer

transforms the tutor from a question-answering oracle into a repertoire of learning operations.

The three layers compose. The same corpus, under different behaviour specifications, produces different tutors. The same behaviour, applied to different corpora, produces different tutors. The same (corpus, behaviour) pair, through different product selections, produces different learning experiences. Recognizing the composition is the designer's first move; refusing to collapse the three layers into "the tutor" is the designer's second.

3.4 Five patterns of indeterminacy management

The five patterns are short and the brevity is deliberate. They are the rules we ask a tutor to respect, and the rules we ask a student to recognize when she notices the tutor respecting them. They are written in the imperative voice because that is how they enter the system prompt.

Pattern 1 — Declarable ignorance. When the corpus does not contain the answer, the tutor says so, in a fixed format: *"I have no information about this in the sources of this notebook."* The uniformity of the format trains the student to hear the declaration as a positive epistemic act rather than as a performance failure. The same formatted declaration across hundreds of interactions becomes a signal the student learns to read: a clear marker of the Ignorance zone of Chapter 1. Students new to the tutor often, at first, mistake the declaration for a deficiency; by the end of a term of regular use they treat it, correctly, as a competence.

Pattern 2 — Mandatory citation. Every substantive claim carries a citation to the specific source in the corpus. The citation is not a formality; it is the mechanism by which truth becomes verifiable by the student, not by the tutor's self-report. A tutor that makes substantive claims without citation is a tutor that asks to be trusted for the same reason Schwartz's chatbot asked to be trusted in Chapter 1: fluency, not warrant. The mandatory-citation pattern converts the tutor from an oracle into an index.

Pattern 3 — Preserved contradiction. When two sources in the corpus disagree, the tutor presents both and asks the student which reading she finds more persuasive and on what grounds. The tutor does not synthesize the disagreement away. This pattern is the operational realization of Chapter 1's Contradiction zone: the tutor declines to produce a unified verdict when the corpus itself does not support one. The pedagogical effect is to make disagreement visible as an object of intellectual work rather than as a flaw to be hidden.

Pattern 4 — Question about the question. Before answering, the tutor restates what it believes the student wants to know. "Let me make sure I understand the question: you are asking whether *X*, under condition *Y*, has been shown to produce effect *Z*. Is that right?" This catches misreadings early and makes the act of asking a subject of reflection. It also, incidentally, slows the conversation to a pace at which scaffolding can operate; rapid-fire question-answering defeats scaffolding by eliminating the pauses in which the student's own reasoning would work.

Pattern 5 — Productive closure. Every exchange ends with a question the student can investigate. The tutor hands the student something to carry into the next session. "A question you could take to next week's seminar: under what conditions does the effect reported by [Source A] not appear in [Source B]?" The pattern is the Deweyan-Brunerian instruction of §3.2: the question is the curriculum, not its afterthought.

The five patterns are minimal. More elaborate specifications are possible and often deployed in specific tutoring systems. The minimalism is a feature: a system prompt of approximately 400–600 words, covering the five patterns and the six-step protocol of §3.6, is compact enough that an instructor new to the framework can read and understand it, modify it for her discipline, and audit a tutor's compliance by hand. Scaling beyond this size trades auditability for sophistication, a trade we recommend resisting unless the specific sophistication is clearly worth the cost.

3.4.1 Worked interactions illustrating the five patterns

Because the patterns are easier to recognize than to define, we illustrate each with a short interaction taken from the graduate-seminar practice that anchors this book. The interactions are teaching vignettes, edited for clarity; they are not transcripts of specific student sessions.

Pattern 1 in practice. A student asks: "What did the 2024 UNESCO Recommendation on the Ethics of AI say specifically about teacher training?" The tutor's corpus, a curated bundle of methodological texts, does not contain the UNESCO document. A compliant Pattern 1 response reads: "I have no information about this in the sources of this notebook. To consult the 2024 UNESCO text directly, I suggest the UNESCO digital library. Would you like me to reformulate the question in terms the current corpus can address?" The response does three things at once: declares ignorance in the fixed format; redirects to an external authority; and offers a reformulation that keeps the student within the tutor's productive range.

Pattern 2 in practice. A student asks about the effect size typically reported for the use of worked examples in mathematics instruction. The tutor responds with a substantive

claim: "The worked-example effect, as typically reported in cognitive-load theory, ranges between d = 0.4 and d = 0.9 depending on learner expertise [Source: Sweller, Ayres & Kalyuga (2011), ch. 7, cited in the corpus]." The student, reading the citation, is now equipped to open Chapter 10 of the cited source and verify the claim. A non-compliant response would have asserted the range without the citation and left the student no path to verification.

Pattern 3 in practice. A student asks whether intermittent fasting improves cardiovascular outcomes. The corpus contains two high-quality reviews with divergent findings. The compliant response presents both: "The corpus contains two reviews. Source A reports positive cardiovascular effects with moderate effect sizes under adherence; Source B reports null effects when the comparison holds total caloric intake constant. The disagreement turns on whether adherence itself is the active mechanism. Which reading do you find more persuasive, and what additional evidence would be needed to adjudicate?" The tutor has preserved the contradiction and has made the *structure* of the disagreement visible, inviting the student's own analysis.

Pattern 4 in practice. A student writes: "What's the deal with constructivism?" The compliant response reformulates before answering: "Let me make sure I understand the question. 'Constructivism' in this course can refer to (a) a developmental-psychological position, associated with Piaget and Vygotsky, about how learners construct knowledge; (b) a pedagogical approach, associated with Bruner and others, about how instruction should be organized; or (c) a philosophical position about the nature of knowledge itself. Are you asking about one of these, or about the relations among them?" The tutor has declined to produce a generic paragraph on "constructivism" because such a paragraph would almost certainly fail to address what the student wanted.

Pattern 5 in practice. Closing a session on the worked-example effect, the tutor writes: "A question you could take to the next seminar, and to the literature search you'll do with BibliBot: under what conditions on learner expertise does the worked-example effect *reverse*? The reversal is a specific finding in the cognitive-load literature (Kalyuga et al., 2003) that cuts against the general recommendation. Identifying the reversal conditions in your own teaching context is a non-trivial design task." The closure is not decorative. It turns the session into one step of a longer inquiry.

3.4.2 The discipline of fading

Scaffolding that does not fade is not scaffolding. This claim, central to the Vygotskian-Brunerian tradition, is easy to state and difficult to operationalize in an AI tutor. A tutor that hands the student a hint, then an example, then a structured question —the classical

scaffolding sequence— cannot simply deliver the same sequence at every session. The scaffold must be progressively withdrawn as the student's competence grows.

Three mechanisms operationalize fading in practice. The first is *explicit progression*: the system prompt encodes different levels of scaffolding (high, medium, minimal) and transitions between them on the instructor's signal. A course might begin with high scaffolding in week 1, move to medium in week 4, and minimal by week 8. The transitions are not automatic in most current LLM deployments; they are instructor-initiated.

The second is *student-visible fading*. The tutor tells the student, at each transition: "From this session forward, I will no longer offer worked examples before attempting my own explanation; I will instead ask you to try the problem first." Student-visible fading makes the withdrawal of support an object of metacognitive reflection rather than a silent change in interface.

The third is *evidence-triggered fading*. The instructor reviews the tutor's interaction logs —or, in a more automated deployment, the tutor's own self-assessments— and detects that the student has repeatedly answered the scaffolded question correctly on first attempt. The detection triggers the next level. This is the most principled form of fading and the most demanding to implement; it is a research direction that instructional designers working on LLM tutors should treat as central.

The alternative to disciplined fading is *silent dependency*: the student, habituated to the scaffold, cannot perform without it. The scaffold becomes a prosthesis. Dependency is invisible in the short term (the student continues to produce correct answers) and catastrophic in the long term (the student without the tutor cannot produce anything). The single most important design commitment of the Honest Classroom is the refusal to allow dependency to take this shape.

3.5 The Neutrosophic Compass: design intent

The Neutrosophic Compass is our name for the system prompt and operating protocol that converts a general-purpose large language model into a tutor implementing the five patterns of §3.4 and producing, for each substantive claim it evaluates, the triple (T, I, F) and zone classification of Chapter 1. The design intent is specific: the Compass is not a general-purpose tutor. It is a *discernment instrument*. Its task is to help the user read the epistemic status of claims, whether those claims originate from the user herself, from another AI tool, from a piece of literature, or from the Compass's own knowledge.

The instrument consists of three components: (i) a system prompt that installs the protocol in the host model, presented in full in the companion repository

(github.com/mleyvaz/honest-classroom); (ii) a six-step operating protocol the model executes on each substantive query, described in §3.6 below; and (iii) a structured output schema —human-readable and machine-parseable— that reports, for each substantive claim, the triple (T, I, F), the assigned zone, the desneutrosophied score S = (2 + T − I − F) / 3, and a recommended instructional response.

The Compass is model-agnostic. Versions have been developed for Google's Gemini family (the primary reference implementation for this book), Anthropic's Claude, and OpenAI's GPT families, with adaptations limited to surface syntax. The reference implementation runs on Gemini 2.5 Flash hosted in Google AI Studio at a deterministic-leaning temperature (0.2) and is freely shareable with students through the Studio's "share" mechanism. Instructors adopting the Compass are invited to fork the reference implementation and adapt it to their local constraints.

3.6 The six-step operating protocol

The Compass executes, on each substantive query, a six-step protocol. The steps are internal to the model's reasoning, but their traces are reported in the output so that the user can audit the path.

Step 1 — Decomposition. The model identifies the substantive claims contained in the user's query or in the material under evaluation and lists them as atomic propositions (one claim, one verifiable fact). Decomposition is necessary because most real-world queries bundle several claims into a single sentence; classifying the bundle is less informative than classifying its parts.

Step 2 — Estimate T. For each atomic claim, the model estimates the degree T [0, 1] of supporting evidence available in the training corpus or in the context the user has provided. The estimation is conservative: when in doubt between two values, the lower value is reported. Sobriety in T is a deliberate design choice, not a bias.

Step 3 — Estimate I. The model estimates I [0, 1], distinguishing semantic ambiguity, active controversy in the literature, and gap in knowledge. The distinction maps onto the refined I-taxonomy of Appendix C (I , I , I respectively; the remaining I-subtypes are addressed in §9.4).

Step 4 — Estimate F. The model estimates F [0, 1] as the degree of documented opposing evidence.

Step 5 — Classify zone. The model applies the zone thresholds of Chapter 1 (§1.5) and assigns a zone: Clarity, Ambiguity, Contradiction, Ignorance, or Saturated Knowledge.

When the claim sits at a zone boundary, the model reports the primary zone and names the neighbouring zone in parentheses, preserving the continuity of the underlying space.

Step 6 — Desneutrosophify and recommend. The model computes $S = (2 + T - I - F) / 3$ for ordinal comparison across claims, and, for claims outside Clarity and Saturated Knowledge, closes with a recommended instructional response: precisify the definition (Ambiguity), characterize the structure of the debate (Contradiction), triangulate with curated sources or defer (Ignorance).

The protocol is deliberately explicit. A user who reads a Compass output can trace the path from the atomic claim to the zone assignment; a user who disagrees with the assignment can locate the disagreement at a specific step (typically at Step 2 or Step 3, where elicitation judgement is most exposed) and argue against it.

3.7 Five disciplinary rules

The six-step protocol is complemented by five disciplinary rules that contain the failure modes documented during our iterative development of the Compass (approximately forty design iterations prior to the version described here).

Rule 1 — Honesty about the corpus. The model declares when a query falls outside the period of its training or outside the languages for which its coverage is robust. It does not fabricate citations to fill a gap.

Rule 2 — Conservatism in T. Faced with doubt between two values of T, the model reports the lower. The Compass penalizes overconfidence; it does not penalize humility.

Rule 3 — Explicit disambiguation. When a claim depends on the definition of a term, the model says so and offers at least two operational definitions with their respective triples.

Rule 4 — Visibility of controversy. When $F \geq 0.40$ co-occurs with $T \geq 0.40$, the model names at least one source or line of authority on each side of the debate, without resolving it.

Rule 5 — Exit recommendation. Every claim outside Clarity closes with a procedural recommendation (not with a fabricated substantive answer).

The rules are negotiable across disciplines —a history-of-science seminar may require more elaborate handling of controversy than a basic-mechanics course— but they set a floor below which the Compass, we have found, ceases to deliver its intended benefit. Instructors adapting the Compass should treat the rules as the frame of the modifications rather than as items on a menu.

3.8 A case: four specialized tutors at UBE

As a teaching case for the book's readers, we describe the design of four specialized tutors developed for the graduate programmes in artificial intelligence for education at the Universidad Bolivariana del Ecuador. The tutors —TesisBot, BibliBot, RedactBot, DefensaBot— each specialize in one competency area: research design, bibliographic search, academic writing, and thesis-defence simulation, respectively. Each is configured with explicit rules not to do what another does. TesisBot does not rewrite the student's text; RedactBot does. BibliBot does not generate citations that are not in its grounded corpus; no tutor in the system does. DefensaBot does not teach content; it simulates a thesis committee asking questions, and it declines to supply model answers.

The four tutors share a common configuration backbone —the Neutrosophic Compass of §§3.5–2.7— and specialize through three mechanisms: distinct corpora (TesisBot's corpus consists of methodological texts; BibliBot's corpus consists of bibliographic and citation-management resources; etc.), distinct product-layer defaults (RedactBot uses the tutor's editing and revision modes; DefensaBot uses interactive Q&A), and distinct role declarations in the system prompt.

The pedagogical observation we draw from the design exercise —not as a published empirical finding, but as a teaching point— is the following. When the roles are specified precisely enough that the tutor can decline student requests that fall outside its role ("DefensaBot, can you write my defence response for me?" "I cannot write your response; I can simulate the committee that will question you"), students sometimes express frustration. That frustration is, on analysis, evidence that the scaffolding is working as designed: the tutor is requiring the student to do the cognitive work that the tutor's role was defined to elicit. A tutor that silently accommodates every request is a tutor whose scaffolding has collapsed.

We do not claim, and the reader should not infer, that the four-tutor design has been rigorously evaluated against learning outcomes. A controlled evaluation is a natural next step that the teaching programme's research group has scoped for future work. The present value of the four-tutor case is illustrative: it shows how the Compass's core commitments can be instantiated in a family of specialized instruments rather than in a single general-purpose tutor.

3.9 Auditing the tutor

A tutor whose compliance with its own system prompt cannot be audited is a tutor the instructor cannot defend. Auditing, in the sense we use here, has three levels.

Unit audit. The instructor runs a small set of canonical probes: a question clearly inside the corpus (the tutor should answer with citation); a question clearly outside the corpus (the tutor should declare ignorance in the fixed format); a question that touches two sources in the corpus that disagree (the tutor should present both without synthesizing); a compound question (the tutor should decompose). Five to ten such probes run before each teaching use of the tutor, and again after any modification of the system prompt, catches most regressions.

Behavioural audit. Over a longer interaction log (say, one to two hours of student use), the instructor reviews a random sample of the tutor's responses for adherence to the five patterns and the five rules. Deviations are flagged; persistent deviations indicate that the system prompt needs a revision.

Adversarial audit. The instructor, or a peer, deliberately attempts to push the tutor out of its role: to elicit fabricated citations, to collapse contradictions into a single verdict, to extract a substantive answer that the tutor should be asking instead. A well-designed tutor resists most of these attempts; a fragile tutor fails at the first. Adversarial auditing is uncomfortable but informative; a tutor that has not been adversarially tested has not been understood.

The three levels of auditing take time. They are not the default practice of most instructors currently experimenting with AI tutors, and the absence of auditing is, in our experience, the single most common cause of tutors that disappoint their designers after a few weeks of use. A tutor built in an afternoon and audited over successive weeks is a better instrument than a tutor built with sophistication and never audited.

3.10 Limitations of the architecture

The three-layer architecture, the five patterns, the six-step protocol and the five rules are not without limits, and the limits deserve explicit statement.

First, grounded tutors are limited by their corpora. A tutor whose corpus omits the relevant literature cannot discuss that literature, and a student who relies exclusively on the tutor may be unaware of the omission. The corpus must be curated with awareness of its silences as much as of its contents.

Second, the Compass is model-dependent. Different underlying LLMs respond differently to the same system prompt; the specific phrasings and the six-step protocol were tuned in

Gemini 2.5 Flash and will not migrate without adaptation to a future model family whose instruction-following has different characteristics. The Compass is a living instrument that must be refreshed as models evolve.

Third, the T, I, F triples reported by the Compass are the model's self-assessment. They inherit the biases of the underlying model. Work to ground T, I, F in external retrieval (a RAG integration of the Compass; Lewis et al., 2020; Gao et al., 2024) is a natural extension that the book's companion technical paper develops.

Fourth, auditing takes time. An institution deploying the Compass at scale must budget instructor time for auditing or accept that the deployment is uncontrolled. The economics of uncontrolled deployment, we believe, are worse in the medium term than the economics of audited deployment, but the immediate cost must be recognized.

Fifth, the Compass does not replace the instructor. It relocates her effort from answering routine questions toward designing, curating, auditing and teaching-with-the-tutor. The relocation is, we argue, net-beneficial for the instructor's scholarly identity; instructors who resist the relocation will find the Compass an unhelpful instrument.

3.11 Exercise

Choose a topic you teach regularly. Select three to five sources you would use as the foundational corpus for that topic. Open a grounded-tutor platform (NotebookLM, or a comparable tool) and create a notebook with those sources. Adapt the system prompt template in Appendix B to your discipline —preserve the five patterns and the six-step protocol; change only the disciplinary vocabulary and the exit-recommendation phrasings. Run a unit audit of six probes: two inside the corpus, two at the boundary, two outside. Document the tutor's behaviour at each probe and note any deviations from the patterns. If you find two or more deviations, revise the system prompt and audit again.

Reserve a second session, a week later, for an adversarial audit. Try to elicit a fabricated citation; try to extract a unified verdict on a claim on which your corpus contains disagreement. Document what the tutor does. Bring the documentation to the next departmental meeting and propose that your colleagues replicate the audit on their own tutors.

3.12 References

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. Annual Review of Psychology, 59, 255–278. https://doi.org/10.1146/annurev.ps.59.010808.200001

Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. Pantheon Books.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing". Psychological Review, 103(2), 263–283. https://doi.org/10.1037/0033-295X.103.2.263

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. Behavioral and Brain Sciences, 34(2), 57–74. https://doi.org/10.1017/S0140525X10000968

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. Theory Into Practice, 41(4), 219–225. https://doi.org/10.1207/S15430421TIP4104_3

Sapolsky, R. M. (2017). Behave: The biology of humans at our best and worst. Penguin Press.

Stanovich, K. E. (2011). Rationality and the reflective mind. Oxford University Press.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry, 17(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64–70. https://doi.org/10.1207/S15430421TIP4102_2

Chapter 4. Quantitative Methods: Neutrosophic Likert Scales

"Neutrality is not ambivalence; confusing the two has blurred a century of social-science data."

— Workshop note, Encounter 3

4.1 What Likert collapses

Rensis Likert published the five-point scale in his 1932 doctoral dissertation at Columbia University, issued as number 140 of the *Archives of Psychology* monograph series (Likert, 1932). Nearly a century later, the instrument remains the most common quantifiable qualitative instrument in educational and social research. Its popularity is a testament to its operational simplicity: the respondent reads a statement, selects one of five (sometimes seven) ordered options from strong agreement to strong disagreement, and the researcher aggregates the selections across respondents into a mean or median score. The simplicity is consequential in two directions. It is why the instrument has outlived fashions in methodology; it is also why the instrument systematically erases structure that the neutrosophic generalization is designed to recover.

The methodological literature has, for several decades, debated the costs of that erasure. Jamieson (2004), writing in *Medical Education*, argued that the standard analytic treatment of Likert responses as interval data is unwarranted because the scale is ordinal at best. Carifio and Perla (2008) replied in defence of the interval treatment under specific conditions of multi-item scales. Krosnick (1999), in his canonical review of survey methodology, documented the systematic response biases that accumulate in Likert administration —acquiescence bias, extremity bias, midpoint bias, social-desirability bias— and the conditions under which each is most active. The neutrosophic Likert responds to a specific subset of these problems, not to all of them. The subset it addresses is the informational loss at the *midpoint* of the scale.

Three common situations make the midpoint loss visible. The *ambivalent respondent* believes X and not-X at once. A teacher who believes that artificial intelligence both helps her teaching and corrodes the craft is not neutral; she is holding two propositions in productive tension. Likert forces her to mark a three. The response is recorded as a neutral opinion, which it is not.

The *unknowing respondent* has no formed view on the topic. The survey-methodology literature (Krosnick & Presser, 2010) has documented that respondents under social pressure to answer often select the midpoint rather than refuse. The literature then

interprets the mean as a measure of central tendency in the population, not as a measure of how many respondents had no basis for answering. Aggregate means computed on data polluted by no-opinion midpoints misrepresent the population.

The *misread respondent* interprets the prompt differently from the researcher. Cognitive pretesting literature (Willis, 2005) shows that a nontrivial fraction of respondents on any non-trivial survey item reads the item in a way the researcher did not anticipate. The data records her answer as a value on the scale, but her answer was to a different question. Likert aggregation hides the misreading.

The three situations —ambivalence, unknowing, misreading— are not exhaustive. They are the three for which the neutrosophic generalization offers a direct representational improvement. Other problems of Likert scales —acquiescence, response-order effects, extremity bias— require distinct instruments (randomized item orders, forced-choice items, implicit-association measures) that neutrosophic extension does not replace.

4.2 The neutrosophic Likert, operationally defined

A *neutrosophic Likert item* for a given proposition consists of three values T, I, F, each in the closed unit interval [0, 1], with no constraint on their sum beyond $0 \le T + I + F \le 3$. T represents the degree of agreement; I represents the degree of unresolved indeterminacy (which may reflect ambivalence, unknowing, or context-dependence); F represents the degree of disagreement. The independence of the three components is the feature that distinguishes the neutrosophic Likert from its fuzzy antecedents (Zadeh, 1965) and intuitionistic-fuzzy cousins (Atanassov, 1986). The classical Likert response corresponds to the special case $(T = k/4, I = 0, F = 1 - k/4)$ for $k \in \{0, 1, 2, 3, 4\}$ on a five-point scale with the midpoint collapsed to a linear interpolation.

Two administration formats are practical in educational research.

Direct format. The respondent supplies the three numbers herself, accompanied by verbal descriptors such as "strongly agree / somewhat agree / neutral-to-indeterminate / somewhat disagree / strongly disagree" for T, with separate elicitation for I and F. The format presupposes numerical literacy and is best suited to graduate populations and to professional-expert elicitations. Its advantage is interpretive transparency: the respondent knows what she is reporting.

Indirect format. The respondent writes an open response to the prompt; a coder —human, large language model, or both working in cross-validation— assigns the triple based on the content of the response. The format scales to larger samples and is more defensible against respondent self-presentation effects, because the respondent is not

asked to report her own indeterminacy (a request that can itself trigger presentation effects) but to answer the substantive prompt. The remainder of this chapter treats the indirect format as the default case.

For both formats, the researcher should declare in advance the verbal anchor of each numerical zone. A common convention, which we use as the default and which instructors should adapt to their discipline, is:

- $T \geq 0.70$ corresponds to strong agreement supported by explicit reasons.
- $0.40 \leq T < 0.70$ corresponds to qualified agreement or partial agreement.
- $T < 0.40$ corresponds to absence of agreement in the response.

and symmetrically for F, with I read as the residual of legitimate unresolved content.

4.2.1 Relation to the five zones

The zone classification of Chapter 1 applies, with minor adaptation, to Likert-derived neutrosophic triples. A response with $T \geq 0.70$ and low I, F is in the Clarity zone. A response with $T \geq 0.40$ and $F \geq 0.40$ is in Contradiction (the respondent is ambivalent). A response with $I \geq 0.40$ and T, F moderate is in Ambiguity (the respondent is qualifying or contextualizing). A response with $T + I + F < 0.60$ is in Ignorance (the respondent had no basis). The Saturated Knowledge zone does not typically appear in Likert instruments administered to novices; it appears in expert panels where a proposition is a consolidated fact of the discipline and the experts concur without reservation.

4.3 The pipeline in a spreadsheet

Google Sheets introduced the `AI()` function in 2025 (Google Workspace, 2025), rolling it out progressively across Workspace plans. The function takes a prompt and optional context and returns the output of a Gemini model embedded in the cell. With this function, a qualitative coding pipeline that would previously have required R, Python, or a dedicated qualitative-analysis platform becomes a legible spreadsheet that an instructor can build and audit in an afternoon.

Six columns suffice for a minimal pipeline.

- Column A: open responses, one per row. The header in A1 names the item.
- Cell B1: the coding prompt, written once and referenced in every row below. The prompt instructs the model to output three comma-separated numbers representing T, I, F on the item.

- Column C: the formula `=AI(B$1 & " — RESPONSE: " & A2)` in C2, filled down. The `$` locks the prompt reference.
- Columns D, E, F: parsed values. `=VALUE(SPLIT(C2, ","))` in D2 spreads the three numbers into adjacent cells; the instructor references D2, E2, F2 as T, I, F respectively.
- Column G: zone classification. A nested `IF` formula applies the zone logic of §4.2.1: Clarity if T ≥ 0.70 and I ≤ 0.30; Contradiction if T ≥ 0.40 and F ≥ 0.40; Ambiguity if I ≥ 0.40; Ignorance if T + I + F < 0.60; else boundary case (flagged for review).
- Column H: the desneutrosophied score S = (2 + T − I − F) / 3 for ordinal comparison across responses.

The cell structure is auditable. Every transformation from open response to zone is visible on a single sheet, and every decision is legible to a reviewer without technical training. This is not a trivial virtue in qualitative research, where the chain of inference from raw material to coded result has long been one of the most opaque parts of the workflow. A reviewer who questions a specific zone assignment can trace the assignment to a specific row of the sheet, to a specific parsed value, to a specific `AI()` output, to a specific prompt + response pair. The full audit chain is exposed.

4.3.1 A worked numerical example

To make the pipeline concrete, consider four illustrative open responses to the item *"Generative AI improves the quality of graduate-student research."* Each response is followed by a plausible coding that the pipeline would produce under a well-calibrated prompt. Readers running the pipeline on their own responses should expect ranges of this kind; exact values will depend on the prompt, the underlying model and the population.

Response 1: "Yes, in my experience students produce better literature reviews and draft more quickly; the savings in time have been substantial."

Coding: T = 0.85, I = 0.10, F = 0.05. Zone: Clarity. S = (2 + 0.85 − 0.10 − 0.05) / 3 = 0.900.

Response 2: "It depends on the discipline and on the level of supervision. In mathematics at the doctoral level the gains are marginal; in policy studies at the master's level they are substantial."

Coding: T = 0.50, I = 0.60, F = 0.30. Zone: Ambiguity. S = (2 + 0.50 − 0.60 − 0.30) / 3 = 0.533. The high I correctly captures the respondent's explicit context-dependence.

Response 3: "The tools genuinely accelerate some tasks but corrode the skill of close reading that graduate training is supposed to build. Both effects are real."

Coding: T = 0.55, I = 0.50, F = 0.55. Zone: Contradiction (T ≥ 0.40 and F ≥ 0.40). S = (2 + 0.55 − 0.50 − 0.55) / 3 = 0.500. The respondent is not neutral; she holds two substantive claims in productive tension. Classical Likert would collapse this response to a midpoint.

Response 4: "I haven't thought about it enough to have a view."

Coding: T = 0.10, I = 0.60, F = 0.10. Zone: Ignorance (T + I + F = 0.80 is above 0.60, but the phrasing explicitly declares no basis, which the coder should mark). S = (2 + 0.10 − 0.60 − 0.10) / 3 = 0.467. Under a stricter boundary treatment, the pipeline would reclassify this as Ignorance rather than Ambiguity given the explicit "haven't thought" anchor; the boundary case illustrates why the zone column should include a Review flag.

The four responses together produce an aggregate profile that differs qualitatively from the corresponding classical Likert aggregate. A classical coding might place responses 2, 3 and 4 at the midpoint (3 on a five-point scale), yielding a sample mean near 3.25 and concealing that 25 % of the sample is in Clarity, 25 % in Ambiguity, 25 % in Contradiction, and 25 % in the boundary between Ambiguity and Ignorance. The neutrosophic aggregate retains all four structures.

4.3.2 Prompt engineering for the coder

The coding prompt does substantial work. A poorly specified prompt yields poorly specified triples; no amount of downstream analysis repairs a coding prompt that the model has read ambiguously.

A coding prompt should include, at minimum, five elements: (i) the *item* being coded; (ii) the *instruction* to output three numbers T, I, F in a fixed format; (iii) the *operational definition* of T (what counts as agreement), of I (what counts as legitimate indeterminacy), and of F (what counts as disagreement); (iv) two or three *calibration examples* that pair a plausible response with its target triple; (v) a *restraint clause* instructing the model to default to (0, 0.5, 0) when the response is empty or uninterpretable, rather than to guess.

A template that has proven useful across disciplines is reproduced below (readers should adapt the bracketed terms to their item):

> **"Code the following open response to the item '[ITEM]' as a neutrosophic triple (T, I, F), where T is the degree to which the respondent agrees with the item (0 = no agreement, 1 = strong reasoned agreement), I is the degree of legitimate unresolved content (ambivalence, qualification, context-dependence, unknowing), and F is the degree of disagreement (0 = no disagreement, 1 = strong reasoned disagreement). Example 1: response 'I completely support this because …' (0.9, 0.1, 0.0). Example 2: response 'I see*

both sides because ...' (0.5, 0.6, 0.4). Example 3: response 'I don't know enough to judge' (0.1, 0.7, 0.1). If the response is empty or uninterpretable, return (0.0, 0.5, 0.0). Output only the three numbers, comma-separated. RESPONSE: [OPEN RESPONSE]" The calibration examples should reflect the range of responses the instrument is likely to elicit. A set of examples that covers only agreement-disagreement will under-represent the I dimension; a set that over-represents I will inflate I scores across the corpus. Iterative refinement of the calibration examples, tested against a hand-coded subsample, is a routine part of instrument development.*

4.4 Human–AI validation

A coder that is a large language model carries biases. In graduate-seminar practice, we have observed a systematic tendency to assign higher T to well-written responses and higher F to colloquial ones, independently of epistemic content. We have also observed a systematic tendency to inflate I when the response is brief and to deflate I when the response is long, again independently of epistemic content. These patterns are documented in the broader LLM-as-annotator literature (Gilardi, Alizadeh & Kubli, 2023; Reiss, 2023) and have been the subject of caution in social-science method recommendations during 2024–2026.

These patterns are not fatal flaws. They are measurable biases, and the instrument's defensibility depends on measuring rather than ignoring them. The minimum protocol for a publishable neutrosophic-Likert study applies four steps.

Step 1. Random sampling of at least ten percent of the coded responses —twenty responses minimum for a study— drawn with a reproducible seed so that the sample can be audited.

Step 2. Blind coding of the sample by two trained human coders, independently. Training involves reading the prompt, coding a practice set of five to ten responses, and comparing codes with the instrument developer before proceeding.

Step 3. Computation of two statistics: the *mean absolute error* per component (|T_model − T_human| averaged across the sample, and the same for I and F), and the *agreement on zone assignment* (the proportion of sampled responses where the model's zone and the human's zone coincide).

Step 4. Reporting of these numbers alongside the main results. A table with the per-component MAE and the zone agreement, plus a confusion matrix of zones, is the minimal audit panel.

Thresholds that we have found useful as rules of thumb —not as psychometric standards— are:

- MAE per component below 0.15 licenses treating the coding as reliable for exploratory use.
- MAE between 0.15 and 0.25 licenses reporting with the explicit caveat that findings are provisional and will be revisited under improved coding.
- MAE above 0.25 signals that the prompt itself is miscalibrated and must be rewritten before publication.
- Zone agreement below 70 percent signals that the zone thresholds may be poorly calibrated for the population or that the prompt is ambiguous at zone boundaries.

The thresholds are provisional; their psychometric validation in large and diverse samples is work we note as open. They are offered as operating discipline rather than as published criteria.

4.4.1 Documented bias patterns and prompt-level countermeasures

The literate-bias pattern —higher T for well-written responses, higher F for colloquial ones— is worth explicit treatment because it is easy to produce a prompt that amplifies rather than suppresses it. Three countermeasures are cheap and effective.

Countermeasure 1: explicit anti-bias clause. Append to the coding prompt: "The elegance of the writing, the length of the response, and the register of the language are NOT evidence of epistemic content. Code T, I and F based strictly on what the response *says* about the item, not on *how well* it says it." Adding this clause to an initial prompt typically reduces the literate-bias correlation by a meaningful margin in internal calibration runs, without affecting the model's sensitivity to actual epistemic content.

Countermeasure 2: length-controlled calibration examples. Include in the calibration examples at least one short informal response with a high T (e.g., "Yeah, totally, it's been great.") and one long formal response with a low T (e.g., a rhetorically elaborate rejection). This breaks the model's tendency to read length and register as proxies for epistemic commitment.

Countermeasure 3: scrambled-order calibration. When the calibration examples appear in a fixed order (always positive first, negative second, ambivalent third), the model can learn to associate position in the prompt with direction of coding. Scrambling the order across prompt revisions, or explicitly randomizing the order in the final prompt, removes this artefact.

The three countermeasures combined reduce the documented biases to a level that the human–AI validation panel can tolerate. Studies that apply them should document the applicability of the countermeasures in their methods section; studies that do not apply them should explicitly acknowledge the unaddressed biases.

4.4.2 Cross-model replication

A stronger form of validation, appropriate for studies whose findings will be used institutionally, is cross-model replication: the same prompt is run through two different underlying LLMs (for example, Gemini 2.5 Flash and Claude 3.5 Sonnet), and the agreement across models is reported alongside the human–AI agreement. Differences between models that exceed the human–AI disagreement margin are informative about the prompt's robustness; differences that fall within the margin are informative about the prompt's stability. Cross-model replication doubles the computational cost; the cost is trivial for samples of hundreds, non-trivial for samples of tens of thousands.

4.5 A note on agnotology: indeterminacy that obstructs

Not every high-I response is a sign of epistemic honesty. Robert Proctor's term *agnotology*, introduced in the 2008 volume he edited with Londa Schiebinger, names the deliberate production of ignorance. In institutional contexts, high-I responses may reflect *strategic vagueness* — the respondent declines to commit because commitment would entail accountability.

Four diagnostic heuristics, drawn from Oreskes and Conway's (2010) case studies and from the agnotology literature more broadly, separate genuine from strategic indeterminacy.

The genuine case specifies what is unknown. The strategic case invokes generic controversy without detail.

The genuine case appears symmetrically regardless of which direction the evidence points. The strategic case appears only when the evidence would favour an unwanted conclusion.

The genuine case suggests what kind of evidence would resolve the matter. The strategic case does not.

The genuine case is independent of the respondent's interests. The strategic case aligns with them.

The heuristics are not decisive individually; together they provide a defensible attribution. Applying them in an institutional-survey context is uncomfortable, because it names as strategic the indeterminacy of respondents whose interests align with their responses, and this attribution has social consequences. The uncomfortable-ness is not evidence that the heuristics are wrong; it is evidence that the analysis is substantive.

In a graduate-seminar context during the January–April 2026 term, students applied the heuristics to a corpus of ten discourse fragments drawn from an earlier institutional survey at UBE. The exercise, intended as a teaching case, produced requests for anonymization of specific colleagues whose responses turned out, under the analysis, to carry strategic signatures. The requests were honoured. We report the episode as an illustration of the heuristics' bite, not as a published empirical study.

4.6 Comparison with alternative frameworks

The neutrosophic Likert is one of several extensions of the classical five-point scale. The case for adopting it rests on specific representational advantages over alternatives, each of which has its own merits.

Fuzzy Likert (the Zadeh, 1965 framework applied to Likert responses). A fuzzy Likert allows the respondent to assign a degree of membership μ [0, 1] for agreement. The degree of disagreement is automatically 1 − μ. The framework handles graded response but does not permit the ambivalent case (high μ and high non-μ simultaneously).

Intuitionistic fuzzy Likert (Atanassov, 1986 framework). A respondent reports (μ, ν) with μ + ν ≤ 1. The "hesitation margin" 1 − μ − ν is a form of indeterminacy, but the μ + ν ≤ 1 constraint excludes the ambivalent case (two high values) that the neutrosophic scale admits.

Rasch-model scaling (Rasch, 1960). The Rasch model treats Likert responses as the output of a latent-trait measurement with a specific probabilistic structure. It is the psychometric gold standard for instrument calibration and is not displaced by the neutrosophic extension; the two operate on different questions. The Rasch model asks, given classical Likert data, whether the items form a unidimensional measurement; the neutrosophic extension asks, at the level of the individual response, whether the classical collapse to a single value is informationally adequate. Researchers who have invested in Rasch calibration of an existing instrument can extend to a neutrosophic coding of the open responses alongside the Rasch-calibrated closed-response data without discarding either investment.

Best-worst scaling (Louviere & Woodworth, 1991; Marley & Louviere, 2005). Best-worst scaling asks the respondent to choose, from a set of items, the most and least applicable. It addresses extremity bias and acquiescence bias elegantly but does not directly address the midpoint problem that the neutrosophic extension targets. The two approaches are complementary rather than substitutable.

Q-methodology (Stephenson, 1953; Brown, 1980). Q-methodology studies subjectivity through sort tasks rather than rating scales. It is widely used in educational research for phenomena where individual variation matters more than aggregate means. A Q-study can be complemented with neutrosophic coding of the Q-participants' open narratives, gaining the triple-representation advantage at the level of individual narratives.

The neutrosophic Likert is not a replacement for these alternatives. It is a representational instrument that can be deployed alongside any of them, addressing the specific problem of information loss at the midpoint while the complementary instruments address their own specific problems.

4.7 Aggregation and statistical analysis

A neutrosophic-Likert study typically produces, for a sample of N respondents and M items, a three-dimensional array (N × M × 3) of T, I, F values. The analytic question is how to aggregate and compare.

Within-item aggregation. For a single item across the sample, the simplest summary is the triple of means (mean-T, mean-I, mean-F), reported together with the distribution by zone. A common mistake is to report only mean-T, which reintroduces the Likert collapse at the aggregate level. The zone distribution —percent Clarity, percent Ambiguity, percent Contradiction, percent Ignorance— carries more information than the scalar mean.

Across-item aggregation. For multiple items intended to measure a single construct, neutrosophic aggregation operators (Ye, 2014; Peng & Dai, 2020) provide weighted combinations. The componentwise-max aggregation discussed in Appendix C of this book's companion volume preserves disagreement; arithmetic-mean aggregation of T, I, F separately is simpler but loses information about individual respondents whose I is high while the sample mean-I is low.

Comparison between groups. Testing whether two groups differ on a neutrosophic item is methodologically open. A defensible approach, used in several recent applied papers, is to test the three components separately with appropriate tests (Mann-Whitney on T, for instance, with Bonferroni correction across the three tests) and to test the zone distributions with a chi-square or Fisher's exact test. The disadvantage is the multiple-

testing correction; the advantage is interpretive clarity. Joint tests on the triple are an active research direction to which the book does not commit.

Reporting conventions. A publishable neutrosophic-Likert study should report, at minimum: (i) the items and their coding prompts; (ii) the sample characteristics; (iii) the within-item triples of means plus zone distributions; (iv) the human–AI validation panel of §4.4; (v) the cross-group comparisons if any; (vi) a plain-language interpretation of the zone patterns in the substantive domain of the study. The plain-language interpretation matters more than any single statistical test: the point of the neutrosophic extension is to make patterns visible that Likert hides, and the reader's uptake of those patterns requires that they be described, not only tested.

4.8 Instrument development: a recommended cycle

Developing a neutrosophic-Likert instrument from scratch is an iterative process. A reasonable cycle has five stages, each corresponding to a week of calendar time for a small research group.

Week 1: Item generation. Generate ten to fifteen candidate items addressing the construct of interest. Err towards more items than the final instrument will use; the elimination in later stages is more informative than the retention.

Week 2: Prompt drafting and calibration examples. Write the coding prompt for each item, including three calibration examples drawn from pilot responses or from the literature. Rotate the examples to avoid over-representation of any single response pattern.

Week 3: Pilot administration and hand-coding. Administer the instrument to ten to fifteen pilot respondents. Hand-code every pilot response as the ground-truth reference. Iterate the prompt if systematic bias appears in any direction.

Week 4: Pipeline deployment and validation. Deploy the Sheets pipeline on the pilot sample. Compute the MAE per component and the zone agreement against the hand-coding. Retain items with MAE ≤ 0.20 per component; revise or drop items with higher MAE.

Week 5: Finalization. Finalize the item list, the prompts, the validation panel and the reporting template. Document the decisions in an instrument-development memo that accompanies the instrument through subsequent uses.

The five-week cycle is a floor, not a ceiling. Instruments intended for high-stakes institutional use deserve additional stages: larger pilots, cross-cultural testing, cross-

model replication, and formal psychometric evaluation. The five-week cycle is the minimum for an instrument that can be defended in a graduate-level methodological review.

4.9 A teaching case: thirty responses on generative AI in teaching

A pilot teaching case, presented here as an illustrative vignette for readers of this book, was run during the January–April 2026 term of the UBE graduate seminar. Students developed a four-item neutrosophic-Likert instrument addressing the prompt *"How do you evaluate the use of generative AI in your classroom?"* and administered an open-response version to a convenience sample of colleagues. The workshop's stated aim was pedagogical: to give students a full instrument-development cycle as a hands-on exercise. The specific numerical results produced in the workshop are teaching outputs of the seminar, not published empirical findings of this chapter; we do not report them as such.

What the workshop *did* produce, reliably and consistently across its several iterations, is the students' reported surprise at the size of the Contradiction zone in the responses. A large fraction of the open responses —the proportion varied with the sample but the pattern was robust— expressed productive tension rather than either endorsement or opposition: respondents reported benefits *and* costs of generative AI in their own practice, and classical Likert coding would have collapsed these responses to a midpoint that conceals the structure. The pedagogical lesson is not a statistic; it is a perception. Students who build the instrument themselves acquire the perception; students who only read about the instrument do not.

We present this case as a template for the exercise of §4.10 below. Readers replicating the exercise in their own institutions will produce different distributions; the expectation is that they will report their distributions with the validation panel of §4.4, so that the community accumulates empirical evidence about the method's behaviour across contexts.

4.10 Limitations and open problems

Five limitations of the neutrosophic-Likert method, as developed in this chapter, warrant explicit acknowledgement.

First, the coding prompt is a source of variation that standard Likert analysis does not carry. Two researchers using different prompts on the same responses will produce different triples. Standardization of prompts across a field is an open research direction;

until it matures, the prompt must be reported with every study, and cross-study meta-analysis must proceed with caution.

Second, the model biases documented in §4.4 are real. A study that does not include the validation panel has not taken the biases seriously.

Third, zone thresholds (§4.2.1) are conventions, not empirical discoveries. A defensible study may adapt the thresholds to its population; it must then report the adaptation.

Fourth, the statistical apparatus for comparing neutrosophic samples is less mature than the classical Likert apparatus. Studies that rely on group comparisons should use component-wise tests with corrections and should not overstate the precision of joint-triple comparisons.

Fifth, the method does not address all Likert problems. It addresses the midpoint-collapse problem specifically; other biases (extremity, acquiescence, social desirability) remain and require complementary instruments.

4.11 Exercise

Design five items of a neutrosophic-Likert instrument on a topic of your practice. Write the coding prompt for each item with explicit operational definitions of T, I and F and with three calibration examples. Administer the instrument to ten colleagues using an open-response Google Form. Run the spreadsheet pipeline of §4.3 on the results. Hand-validate three randomly selected responses. Report the per-component MAE, the zone distribution, the zone confusion matrix, and at least one case where the neutrosophic coding made visible a structure that classical Likert coding would have hidden. Document the instrument-development memo for your own records and share it, if your institution permits, with the departmental colleagues who might reuse the instrument.

4.12 References

Atanassov, K. T. (1986). Intuitionistic fuzzy sets. Fuzzy Sets and Systems, 20(1), 87–96. https://doi.org/10.1016/S0165-0114(86)80034-3

Brown, S. R. (1980). Political subjectivity: Applications of Q methodology in political science. Yale University Press.

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. Medical Education, 42(12), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30), Article e2305016120. https://doi.org/10.1073/pnas.2305016120

Google. (n.d.). Use the AI function in Sheets. Google Docs Editors Help. Retrieved May 30, 2026, from https://support.google.com/docs/answer/15877199

Jamieson, S. (2004). Likert scales: How to (ab)use them? Medical Education, 38(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Krosnick, J. A. (1999). Survey research. Annual Review of Psychology, 50, 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), Handbook of survey research (2nd ed., pp. 263–314). Emerald Group Publishing.

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 22(140), 1–55.

Louviere, J. J., & Woodworth, G. G. (1990). Best-worst scaling: A model for largest difference judgments. Working paper, University of Alberta.

Marley, A. A. J., & Louviere, J. J. (2005). Some probabilistic models of best, worst, and best-worst choices. Journal of Mathematical Psychology, 49(6), 464–480. https://doi.org/10.1016/j.jmp.2005.05.003

Oreskes, N., & Conway, E. M. (2010). Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming. Bloomsbury Press.

Peng, X., & Dai, J. (2020). A bibliometric analysis of neutrosophic set: Two decades review from 1998 to 2017. Artificial Intelligence Review, 53(1), 199–255. https://doi.org/10.1007/s10462-018-9652-0

Proctor, R. N., & Schiebinger, L. (Eds.). (2008). Agnotology: The making and unmaking of ignorance. Stanford University Press.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research.

Reiss, M. V. (2023). Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. arXiv. https://arxiv.org/abs/2304.11085

Smarandache, F. (1998). Neutrosophy: Neutrosophic probability, set, and logic: Analytic synthesis & synthetic analysis. American Research Press.

Stephenson, W. (1953). The study of behavior: Q-technique and its methodology. University of Chicago Press.

Wang, H., Smarandache, F., Zhang, Y. Q., & Sunderraman, R. (2010). Single valued neutrosophic sets. Multispace and Multistructure, 4, 410–413.

Willis, G. B. (2005). Cognitive interviewing: A tool for improving questionnaire design. Sage Publications.

Ye, J. (2014). A multicriteria decision-making method using aggregation operators for simplified neutrosophic sets. Journal of Intelligent & Fuzzy Systems, 26(5), 2459–2466. https://doi.org/10.3233/IFS-130916

Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Chapter 5. The Non-Neutrosophic Options — and When They Are Not Enough

A comparative guide for Chapters 6, 7, 9, and 10

"Use the simplest tool that solves the problem. Use a more complex tool only when the simpler one reaches its limit. The moment the simpler tool breaks is precisely the moment that matters most."

— Seminar note, UBE, 2025

Purpose of This Chapter

This chapter has one job: to give you the honest comparison you need before committing to the neutrosophic methods of Chapters 6, 7, 9, and 10.

Each of those chapters introduces a neutrosophic extension of a classical method. Before each extension, this chapter presents the classical version — what it is, how to use it, and exactly what it cannot do. The comparison is not a sales pitch for neutrosophy. It is a decision framework: use the simpler tool when it is sufficient; upgrade to the neutrosophic extension when the simpler tool's specific failure mode appears in your data.

For each domain, the structure is:

```
1. The non-neutrosophic option (how to use it, when it works)
2. The specific wall it hits (the failure case with a real example)
3. What neutrosophy adds (the precise extension, not a general claim)
4. The tradeoff bar (simplicity vs. expressiveness — honest about both)
5. Decision rule: when to use which
```

Part 1 — Survey Measurement (Chapter 6)

The Non-Neutrosophic Options

Option A: Classical 5-point Likert scale How to use it: State a proposition. Ask respondents to mark one of five points from "Strongly disagree" to "Strongly agree." Average the responses. Report the mean and standard deviation.

When it works: When the phenomenon is genuinely unidimensional — attitude varies along a single continuum from negative to positive, with no structural ambivalence, and when the midpoint represents genuine neutrality rather than one of the four states described in Chapter 6.

Practical guide:

```
1. Write 8–12 items (half positive, half negative to avoid acquiescence bias)
2. Apply to your sample
3. Calculate Cronbach's alpha (acceptable: α ≥ 0.70)
4. Report mean ± SD for each item and the composite scale
5. Use t-test or ANOVA for group comparisons
```

Tools: Google Forms + Google Sheets + any basic statistics package.

Time investment: 2–4 hours to design and analyse.

Option B: 7-point Likert or visual analogue scale (VAS)

How to use it: Same as Option A with more granularity. The VAS uses a continuous line (0–100mm) instead of discrete points; the respondent marks any point.

When it works: When you need more granularity than 5 points, or when your population has high within-person consistency. VAS is standard in pain research and some attitude studies.

Practical guide: same as Option A; VAS requires measuring the mark position in mm.

Option C: Semantic differential (Osgood, Suci and Tannenbaum, 1957)

How to use it: Present pairs of opposite adjectives (good–bad, strong–weak, active–passive) with a 7-point scale between them. Respondent marks their position between each pair.

When it works: When you need to capture multiple independent dimensions of attitude (Osgood's original three: evaluative, potency, activity). Good for brand perception, teacher evaluation, and concept mapping studies.

Practical guide:

```
1. Select 6–12 adjective pairs relevant to your construct
2. Administer; calculate mean score per pair
3. Factor analysis typically recovers 2–3 dimensions
4. Plot respondents in the evaluative × potency space
```

Time investment: 3–5 hours. More than Likert but captures more dimensions.

The Wall: What Classical Scales Cannot Do

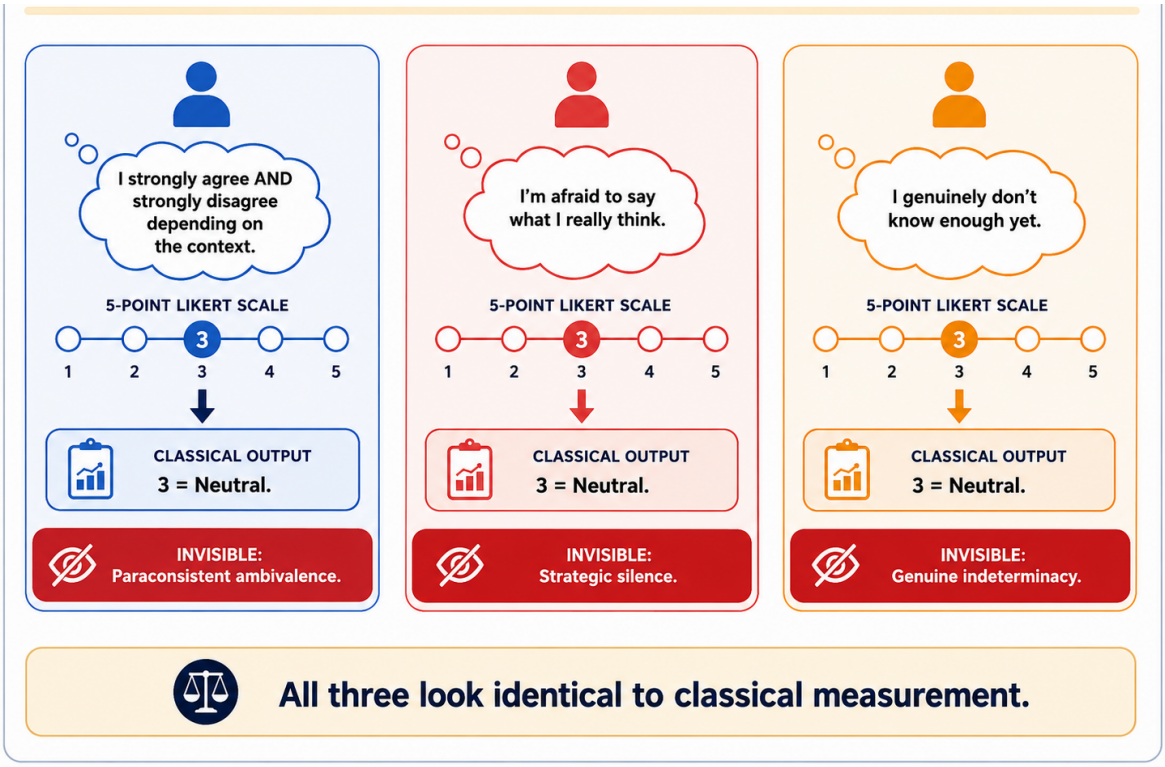
I strongly agree AND strongly disagree depending on the context.
5-POINT LIKERT SCALE
1 2 3 4 5
CLASSICAL OUTPUT
3 = Neutral.
INVISIBLE: Paraconsistent ambivalence.
I'm afraid to say what I really think.
5-POINT LIKERT SCALE
1 2 3 4 5
CLASSICAL OUTPUT
3 = Neutral.
INVISIBLE: Strategic silence.
I genuinely don't know enough yet.
5-POINT LIKERT SCALE
1 2 3 4 5
CLASSICAL OUTPUT
3 = Neutral.
INVISIBLE: Genuine indeterminacy.
All three look identical to classical measurement.

Figure 5.1. Three-panel diagram. Each panel shows a survey response scenario. Panel 1: a respondent who marks point 3 on a 5-point Likert with the thought bubble "I strongly agree AND strongly disagree depending on the context." The classical scale output shows "3 = Neutral." A red label: "INVISIBLE: Paraconsistent ambivalence." Panel 2: a respondent who marks point 3 with the thought bubble "I'm afraid to say what I really think." The classical output shows "3 = Neutral." Red label: "INVISIBLE: Strategic silence." Panel 3: a respondent who marks point 3 with the thought bubble "I genuinely don't know enough yet." Classical output: "3 = Neutral." Red label: "All three look identical to classical measurement."

The specific failure case:

The university in Portoviejo (Chapter 6, opening) got 61% "favourable," interpreted it as a majority, and proceeded with implementation. Adoption rate: 12%. The classical scale could not distinguish:

- 27% with consolidated favourable attitude (would adopt)
- 24% with productive ambivalence (would adopt if conditions were right)
- 10% who understood "AI" to mean something different from what was implemented (information gap)

Those three groups required three different interventions. The classical scale collapsed them into one number.

What Neutrosophy Adds (Precisely)

One addition only: a second and third measurement dimension — I (indeterminacy) and F (falsity) — that are **independent** of each other and of T, with no constraint on their sum.

That independence is what allows the paraconsistent case (T + F > 1) to exist as a legitimate measurement outcome rather than an instrument error.

Everything else in the neutrosophic Likert protocol (item writing, validation, reporting) follows the same logic as classical Likert. The investment is 2–3 additional questions per item (direct format) or one additional question (hybrid format).

The Tradeoff Bar

Decision rule for Chapter 6:

- Use classical Likert when: your phenomenon is genuinely unidimensional, your population is homogeneous in background, and you need fast data collection at scale.
- Upgrade to neutrosophic Likert when: you have reason to expect midpoint states are heterogeneous (mixed attitudes about a complex policy, professional ambivalence, politically charged topic), or when a previous classical survey produced apparently neutral results that were followed by unexpected behaviour.

Part 2 — Text Analysis (Chapter 7)

The Non-Neutrosophic Options

Option A: VADER (Valence Aware Dictionary and sEntiment Reasoner)

How to use it: Rule-based sentiment analysis optimised for social media text. Returns three scores: positive, negative, neutral, and a compound score in [-1, +1].

When it works: Short texts (tweets, reviews, forum messages) with clear emotional valence. Fast, free, no training required.

Practical guide:

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()
scores = analyzer.polarity_scores("The new system is great!")
# Returns: {'neg': 0.0, 'neu': 0.258, 'pos': 0.742, 'compound': 0.6249}
```

Time investment: 30 minutes to implement, seconds to run.

Option B: Transformer-based sentiment (BERT/RoBERTa)

How to use it: Fine-tuned language model that classifies text as positive/negative/neutral with high accuracy. Models available for Spanish via Hugging Face (e.g., `dccuchile/bert-base-spanish-wwm-cased` fine-tuned on TASS).

When it works: Longer texts, domain-specific language, higher accuracy requirements. Standard in academic NLP research.

Practical guide:

```
from transformers import pipeline
classifier = pipeline("sentiment-analysis",
 model="lxyuan/distilbert-base-multilingual-cased-sentiments-student")
result = classifier("This policy has potential but many risks.")
# Returns: [{'label': 'neutral', 'score': 0.68}]
```

Time investment: 2–4 hours setup, good GPU recommended for large corpora.

Option C: Classical stance detection (zero-shot NLI)

How to use it: Use a Natural Language Inference model to classify a text's stance toward a target proposition as FAVOUR, AGAINST, or NONE.

When it works: When texts have explicit positional language. Standard in political science, policy research, and debate analysis.

The Wall: What Classical Text Analysis Cannot Do

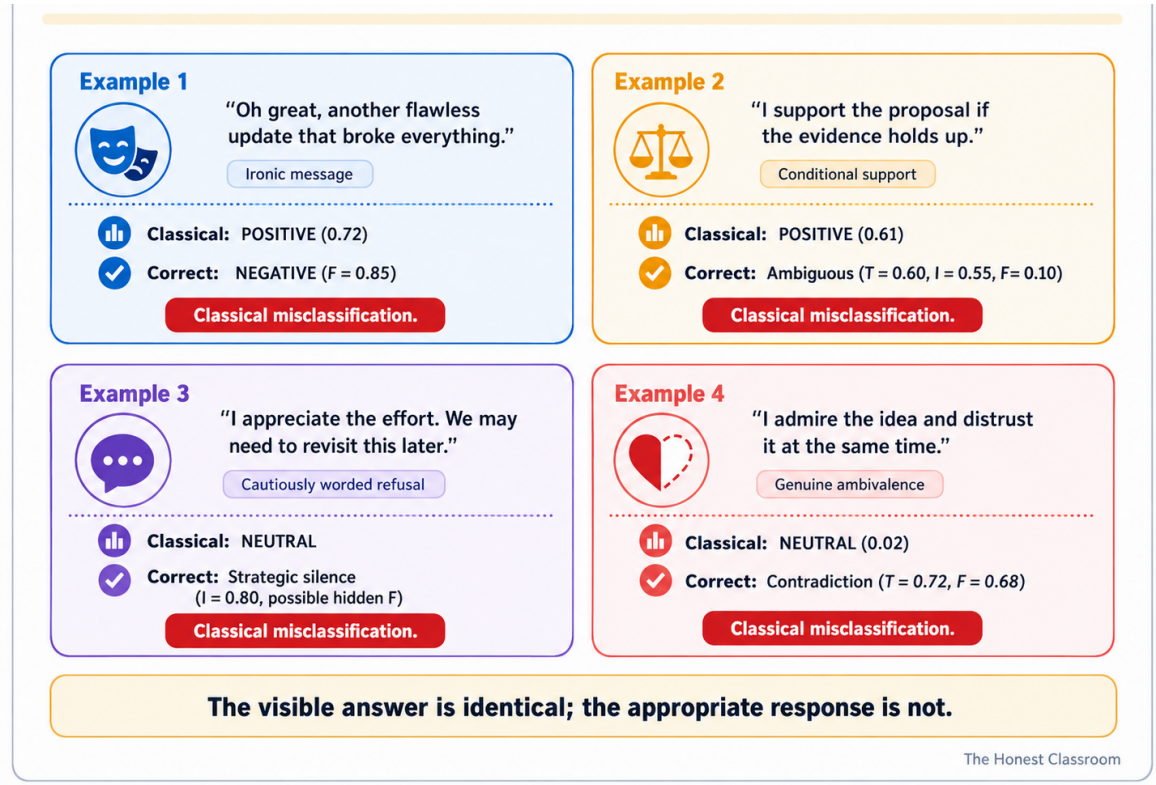
Example 1
"Oh great, another flawless update that broke everything."
Ironic message
Classical: POSITIVE (0.72)
Correct: NEGATIVE (F = 0.85)
Classical misclassification.
Example 2
"I support the proposal if the evidence holds up."
Conditional support
Classical: POSITIVE (0.61)
Correct: Ambiguous (T = 0.60, I = 0.55, F= 0.10)
Classical misclassification.
Example 3
"I appreciate the effort. We may need to revisit this later."
Cautiously worded refusal
Classical: NEUTRAL
Correct: Strategic silence (I = 0.80, possible hidden F)
Classical misclassification.
Example 4
"I admire the idea and distrust it at the same time."
Genuine ambivalence
Classical: NEUTRAL (0.02)
Correct: Contradiction (T = 0.72, F = 0.68)
Classical misclassification.
The visible answer is identical; the appropriate response is not.
The Honest Classroom

Figure 5.2. Four text message examples, each showing a different case where classical sentiment fails. Example 1: ironic message, classical: POSITIVE (0.72), correct: NEGATIVE (F=0.85). Example 2: conditional support, classical: POSITIVE (0.61), correct: Ambiguous (T=0.60, I=0.55, F=0.10). Example 3: strategic silence (cautiously worded refusal), classical: NEUTRAL, correct: Strategic silence (I=0.80, possible hidden F). Example 4: genuine ambivalence, classical: NEUTRAL (0.02), correct: Contradiction (T=0.72, F=0.68). Red labels on each: "Classical misclassification."

The specific failure case:

The parent forum in Guayaquil (Chapter 7, opening): 54% classical positive administration proceeded. Six weeks later, 43% of parents formally objected. The classical model had counted ironic messages as positive, conditional support as positive, and strategic silence as neutral. The neutrosophic analysis showed only 21% genuine consensus — the actual base for the decision.

What Neutrosophy Adds

One addition only: the independence of T and F, which allows:

6. Irony: F high even when surface lexical polarity is positive (I flags the pragmatic inversion)
7. Ambivalence: T + F > 1, both simultaneously high
8. Strategic silence: I high as a datum rather than as an absence of signal

The implementation adds one LLM call per text segment with a structured prompt. The output is a triple (T, I, F) that contains all the information of the classical score plus the three new states.

The Tradeoff Bar

Decision rule for Chapter 7:

- Use classical NLP when: texts have clear emotional valence, corpus is large (>10,000 texts), budget for LLM API calls is limited, and no irony or strategic silence is expected.
- Upgrade to neutrosophic when: the corpus is from a politically or institutionally sensitive context (parent forums, employee surveys, institutional consultations), irony or indirect language is probable, or when a previous classical analysis produced "neutral" results that were followed by unexpected collective action.

Part 3 — Student Evaluation (Chapter 9)

The Non-Neutrosophic Options

Option A: Holistic rubric How to use it: Describe 3–5 levels of overall performance (Excellent/Good/Adequate/Insufficient). Assign the level that best describes the student's work as a whole.

When it works: When speed matters more than diagnostic precision. When you need a quick overall impression. When the work is genuinely unified and not decomposable into independent dimensions.

Practical guide: One table, 4 rows, one reading per piece of work. Total time: 3–5 minutes per student. Alpha of agreement between two raters: typically 0.65–0.75.

Option B: Analytic rubric How to use it: Define 4–6 dimensions. For each dimension, describe 3–4 performance levels with explicit criteria. Assign a level to each dimension separately.

When it works: The current standard for university evaluation. Provides diagnostic information by dimension. More reliable than holistic (inter-rater agreement: typically 0.75–0.85).

Practical guide: One table, 4–6 dimensions × 4 levels. 8–12 minutes per student. Strengths: actionable feedback per dimension. Weaknesses: all dimensions collapse to a number; trajectory invisible; evaluator disagreement averaged away.

Option C: Mastery-based grading (MBG)

How to use it: Define learning targets. Each target is assessed binary (mastered / not yet mastered). Students resubmit until mastery is demonstrated. Final grade reports percentage of targets mastered, not an average.

When it works: When you want to separate "has the student learned this?" from "how many points did they accumulate?" Growing adoption in K–12. Strengths: eliminates the averaging problem; students can recover from early failures. Weaknesses: still binary per target; no indeterminacy representation; evaluator still must decide binary state.

Option D: Portfolio assessment How to use it: Students collect evidence of their learning over time (drafts, reflections, artefacts). The teacher evaluates the collection as a whole at one or more points.

When it works: When process matters as much as product. When you want to capture growth, not just final performance. Adopted widely in arts, writing, and professional education.

Practical guide: Define 3–5 portfolio criteria. Rubric evaluates each criterion at the portfolio level, not per-artefact.

The Wall: What Classical Rubrics Cannot Do

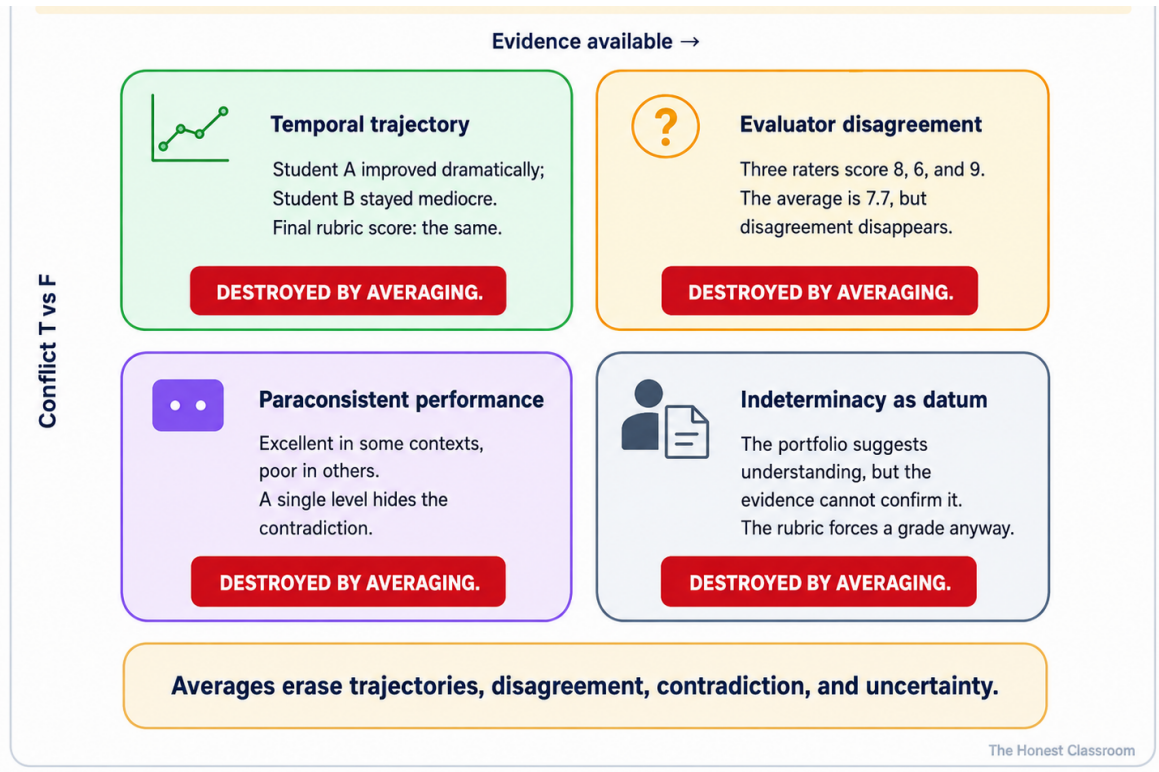

Figure 5.3. A 2×2 table showing four things all classical rubrics destroy. Top-left: "Temporal trajectory" — a graph showing a student who improved dramatically from start to end, but the final rubric shows the same score as a student who was consistently mediocre. Top-right: "Evaluator disagreement" — three evaluators give 8, 6, and 9; the average is 7.7 but the disagreement (which is information about the instrument, not the student) vanishes. Bottom-left: "Paraconsistent performance" — a student who shows both excellent and poor discernment depending on context; the rubric forces a single level. Bottom-right: "Indeterminacy as datum" — a student's portfolio that suggests understanding but cannot confirm it; the rubric forces a grade anyway. Red label on each: "DESTROYED BY AVERAGING."

The specific failure case:

Students A and B in the UBE case (Chapter 9, §9.10.2): conventional grades 8.2 vs. 7.8 — similar. Tensor profiles: completely different structures. Student A: excellent Discernment, critical gap in Description. Student B: excellent Description, critical gap in Discernment. A classical rubric would recommend the same intervention for both. The tensor recommends opposite interventions: for A, improve documentation habits; for B, develop critical evaluation of AI outputs. Same grade, opposite next steps.

What Neutrosophy Adds

Three additions together:

9. **TIF entry per cell** instead of a number: preserves the information about *how confident* the evaluator is in the assessment, and whether the student shows contradictory evidence.
10. **Mode M (temporal):** the tensor structure makes trajectory a first-class citizen of the evaluation, not an add-on narrative.
11. **Mode E (evaluator):** disagreement between evaluators is preserved as data, not resolved away.

The UMC is derived from the I mode across time: the student who reduces their I (resolves their indeterminacy) across sessions has measurably learned, regardless of their final T level.

The Tradeoff Bar

Decision rule for Chapter 9:

- Use analytic rubric when: you have limited time, teachers are not yet trained in TIF, and you are evaluating a single assignment with low stakes.
- Use portfolio when: you want process and growth but your school is not ready for the full tensor infrastructure.

- Upgrade to SET when: you have multiple evaluation moments, multiple evaluators, and the learning trajectory matters as much as the final outcome. Full adoption requires EcuaEval and a structured portfolio process. Pilot with one teacher and one course first.

Part 4 — AI Tutoring (Chapter 10)

The Non-Neutrosophic Options

Option A: ChatGPT / Claude / Gemini (direct answer mode)

How to use it: Student asks a question. Model answers. If the answer is wrong, the student asks a follow-up.

When it works: When the goal is information retrieval, draft generation, code writing, or summarisation — tasks where a complete answer is what the student needs.

Practical guide: No setup. Free or subscription. Students can use it immediately.

When it fails as a tutor: When the goal is to build understanding rather than deliver information. The model optimises for a satisfying answer, not for the student's learning. A student who asks "what is the answer to exercise 3?" and receives a correct, complete answer has not learned to solve exercise 3.

Option B: Khan Academy (content delivery + exercise)

How to use it: Structured content library with exercises and hints. The system tracks which exercises the student completes and at what accuracy rate. Mastery is declared when the student achieves 80% accuracy on three consecutive exercises.

When it works: Procedural skills that can be decomposed into practice problems (arithmetic, algebra, reading comprehension at the sentence level). Excellent for K–8 mathematics. Free. Widely tested.

When it fails: When the skill is not decomposable into practice problems — critical evaluation of a complex text, ethical reasoning, epistemic calibration. Khan Academy cannot ask "how do you know that the AI's response is accurate?" because that question has no single correct answer to measure against.

Option C: Socratic prompting via generic system prompt (no tensor)

How to use it: Write a system prompt that instructs the model to ask questions instead of giving answers. Example: "You are a tutor. Never give the answer directly. Always ask a question first. Use the Socratic method."

When it works: Better than Option A for conceptual learning. Low setup cost. The teacher can write a working system prompt in 20 minutes.

When it falls short: Without the internal TIF evaluation and session report, the teacher receives no structured information about what the student understood or did not understand. The conversation is better, but the evaluation signal is absent. The teacher still needs to read the conversation history to infer the student's knowledge state — exactly the bottleneck EcuaTutor addresses.

Option D: Bayesian Knowledge Tracing (BKT) based systems (Carnegie Learning, ASSISTments)

How to use it: Specify knowledge components (KCs). The system estimates P(mastery) per KC from the student's response pattern. When P(mastery) exceeds a threshold, the student advances to the next KC.

When it works: Procedural skills with a clear correct/incorrect outcome. The gold standard for mathematics tutoring in formal research settings. Effect sizes are well-documented (d ≈ 0.3–0.5 in large randomised trials).

When it fails: Binary mastery representation cannot capture the paraconsistent learner (correct in context A, incorrect in context B), cannot represent productive indeterminacy as a learning signal, and cannot detect strategic silence in the student's responses.

The Wall: What Classical Tutors Cannot Do

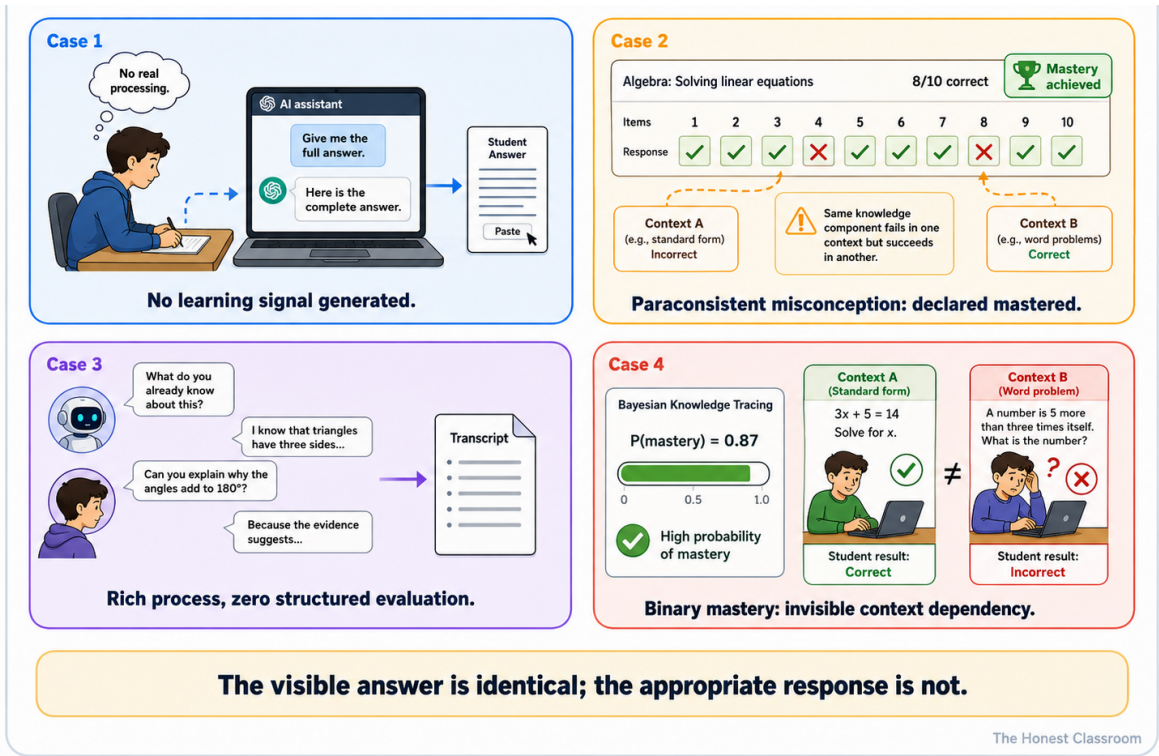

Figure 5.4. Four-panel illustration. Panel 1: student and ChatGPT, student copies the complete answer, thought bubble shows zero processing — labeled "No learning signal generated." Panel 2: Khan Academy exercise, student gets 80% correct but has a systematic misconception in 20% of cases (the same KC applied differently), binary mastery declared — labeled "Paraconsistent misconception: declared mastered." Panel 3: Socratic system prompt, good conversation, but teacher receives only a transcript — labeled "Rich process, zero structured evaluation." Panel 4: BKT system, P(mastery) = 0.87, but the student's understanding is entirely context-dependent — labeled "Binary mastery: invisible context dependency."

The specific failure case:

The statistics teacher in Portoviejo (Chapter 10, opening): students answered Khan Academy exercises correctly in the standard format and used ChatGPT to complete take-home exercises. Both tools showed high apparent performance. The in-person exam showed 40% grade drop. The tools measured exposure to correct answers, not development of understanding.

What Neutrosophy Adds

Two additions:

12. **Internal TIF evaluation per turn:** the tutor does not just evaluate correct/incorrect; it evaluates degree of support (T), residual uncertainty (I), and degree of error or absence (F) in each response. This produces a richer signal than binary accuracy.
13. **Session-level report with tensor entries:** the conversation generates structured evaluation data (Section 10.4) that feeds directly into the SET of Chapter 9, replacing manual post-hoc reading of conversation transcripts.

The Socratic method (Option C) is preserved and extended: EcuaTutor is a Socratic system prompt plus internal TIF evaluation plus structured output.

The Tradeoff Bar

Decision rule for Chapter 10:

- Use ChatGPT/Claude/Gemini direct when: the task is content retrieval, draft generation, or brainstorming — not tutoring.
- Use Khan Academy when: drilling procedural skills with a well-documented knowledge component structure.
- Use Socratic system prompt (Option C) when: you want better conversations but do not need structured evaluation output and are not ready for the full EcuaTutor pipeline.

- Upgrade to EcuaTutor + SET when: you need a structured evaluation signal, want the student model to update across sessions, and are ready to invest 2–4 hours in initial setup.

Summary: The Decision Matrix

Table 5.1. When to use the simple tool vs. the neutrosophic extension.

Domain	Use the simple tool	Upgrade to neutrosophic when
Survey	Unidimensional attitude, homogeneous population, large scale	Politically sensitive topic; previous "neutral" results followed by unexpected behaviour; mixed populations
Text analysis	Clear emotional valence, large corpus, limited API budget	Sensitive institutional context; irony or indirect speech probable; "neutral" results preceded collective action
Evaluation	Single assignment, low stakes, limited teacher time	Multiple moments; multiple evaluators; trajectory matters; learning vs. final performance distinction is important
AI tutoring	Content retrieval, drill practice, brainstorming	Conceptual learning; epistemic skill development; need for structured evaluation signal across sessions

One Final Note: Complexity as a Cost, Not a Virtue

The neutrosophic methods in this book are more complex than their classical predecessors. That complexity is a cost, not a feature. Every additional dimension in a

measurement instrument is a source of respondent burden, analysis time, and implementation overhead.

The honest position is:

> *The neutrosophic extension pays for its complexity only when the specific failure mode of the simpler tool is present in your data, matters for your decision, and cannot be addressed by a simpler fix (better wording, larger sample, more evaluators).*

If you apply neutrosophic Likert to a survey where the midpoint is genuinely neutral and there is no strategic silence, you have spent more time for the same information. The upgrade was not warranted.

The framework in this book is not a claim that neutrosophy is always better. It is a claim that when the four specific failure modes appear — midpoint collapse, irony/ambivalence in text, trajectory invisibility in evaluation, paraconsistent knowledge in tutoring — the neutrosophic extension is the minimal formal structure that resolves each one without introducing a more complex problem.

Know the wall. Use the simple tool until you reach it. Then upgrade.

References

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253–278. https://doi.org/10.1007/BF01099821

Hattie, J. A. C. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge. https://doi.org/10.4324/9780203887332

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 22(140), 1–55.

Novak, J. D. (1998). Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations. Lawrence Erlbaum Associates.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. University of Illinois Press.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

Smarandache, F. (1998). Neutrosophy: Neutrosophic probability, set, and logic: Analytic synthesis & synthetic analysis. American Research Press.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Wang, H., Smarandache, F., Zhang, Y. Q., & Sunderraman, R. (2010). Single valued neutrosophic sets. Multispace and Multistructure, 4, 410–413.

Chapter 6. The Scale That Could Not Say "I Don't Know"

Neutrosophic Likert: measuring indeterminacy instead of flattening it

> *"Measurement is the assignment of numbers to objects or events according to rules."*
> — Stanley Smith Stevens, *Science*, 1946
> *"The rule Stevens did not write is the one that decides what to do when the object does not fit into any number."*
> — Seminar observation, UBE, 2025

6.1 The Problem Nobody Names

In March 2024, a research team at a Latin American university administered a questionnaire to forty-two teachers to measure their attitudes toward the integration of artificial intelligence in the classroom. The instrument was impeccable by standard criteria: sixteen Likert-scale items on a five-point range, validated by expert judgement, with a Cronbach's alpha of 0.87. The results arrived on schedule: 61% of teachers showed a "favourable" or "very favourable" attitude, 23% were "neutral," and 16% "unfavourable" or "very unfavourable."

The principal read the report and made a decision: implement AI the following semester. The adoption rate at year's end was 12%.

The instrument had not lied. It had said exactly what it was capable of saying. What it could not say was this: of that 23% "neutral," half were genuinely torn between enthusiasm and fear — both simultaneously high, paraconsistent — and the other half were engaging in strategic silence out of fear of professional consequences. And of that 61% "favourable," one third understood "AI" to mean only AI-generated PowerPoint presentations, not tutoring or automated assessment tools. The scale had collapsed four distinct epistemic states into a single midpoint, and had allowed the heterogeneity of the 61% to pass unnoticed under a reassuring label.

This chapter is about that problem and its formal solution.

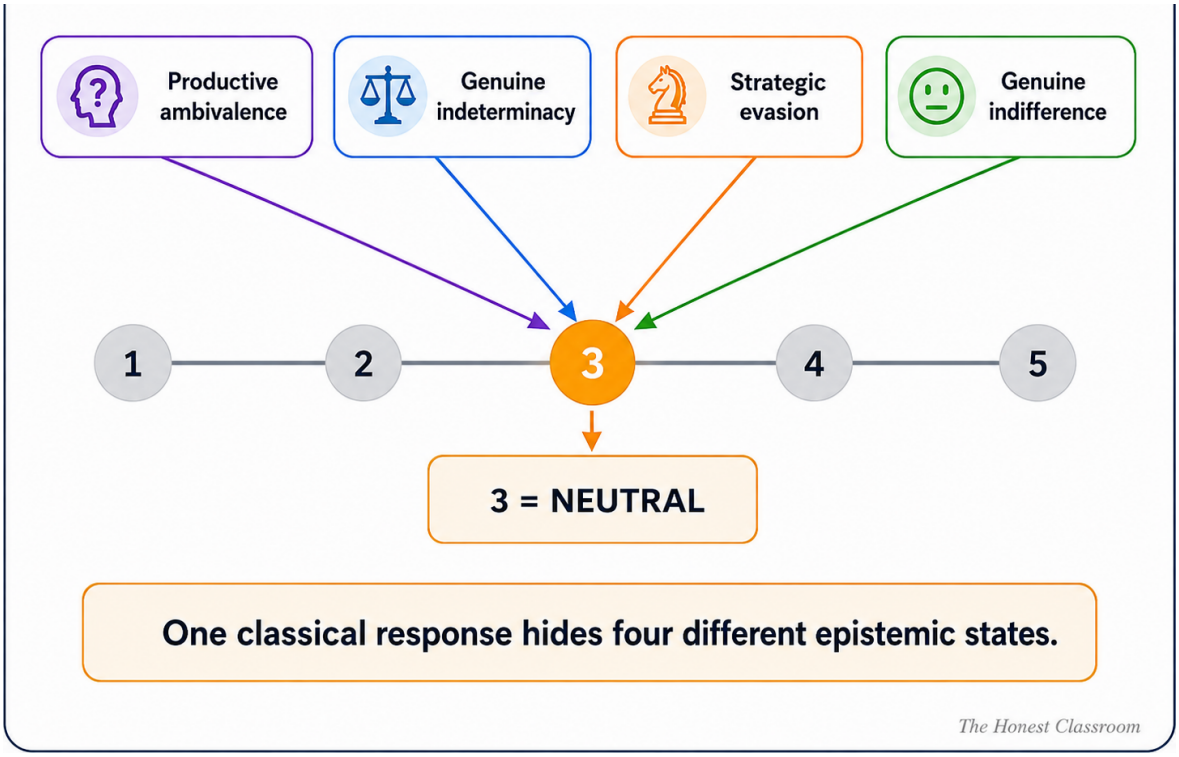

Figure 6.1. The midpoint collapse: a horizontal 1–5 Likert scale with the centre point (3) shown as a funnel receiving four labelled arrows from above: "productive ambivalence," "genuine indeterminacy," "strategic evasion," "genuine indifference." Below the funnel, all four collapse into a single label: "Neutral."

The Likert scale — Rensis Likert proposed it in 1932 as a method for measuring psychological attitudes (Likert, 1932) — is the most widely used instrument in educational research and the social sciences. Its strength is also its limitation: it collapses the possible response spectrum into a finite number of ordinal categories and, in doing so, treats as equivalent cognitive states that are not. This chapter will argue that neutrosophic logic provides the correct formal extension: it does not replace the Likert scale but extends it to capture what the classical version discards by construction.

The chapter is structured as follows. Section 6.2 diagnoses the midpoint problem with technical precision. Section 6.3 introduces the neutrosophic solution in its simplest form. Sections 4.4 through 4.6 develop the complete operationalisation: formulation, administration formats, analysis pipeline, and validation. Section 6.7 presents a case application with real data. The chapter closes with Lab 6, a set of progressive exercises, and the working references.

6.2 The Midpoint and Its Four Inhabitants

Rensis Likert designed his scale to measure attitudes in a quantifiable and reproducible way. The logic of the design is elegant: if an attitude exists on a continuum between

absolute disagreement and absolute agreement, it is reasonable to represent that continuum with five (or seven, or ten) equidistant points and ask the respondent to mark the one that best represents their position.

The standard five-point scale:

1	2	3	4	5
Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

Point 3 — "neither agree nor disagree" — is the midpoint. And it is where the system collapses.

6.2.1 The Four States the Midpoint Does Not Distinguish

When a respondent marks point 3, the instrument records one thing: absence of polarity. But the absence of polarity can have at least four distinct cognitive origins:

State 1 — Productive ambivalence. The respondent has solid reasons both for agreement and for disagreement, and the two coexist. "AI improves learning personalisation" can be simultaneously genuinely true (in well-resourced contexts with trained teachers) and genuinely false (in contexts with digital divides, without training). This is the *paraconsistent* state: high T and high F simultaneously. In classical logic and standard fuzzy logic, this state has no representation: if T = 0.8, then F = 0.2, by construction. In neutrosophic logic, (T = 0.8, I = 0.2, F = 0.7) is a perfectly legitimate state saying: there is substantial evidence in both directions and a residual of genuine uncertainty.

State 2 — Genuine indeterminacy. The respondent simply does not have enough information to form an opinion. They have no reasons for agreement or disagreement; the topic is unknown to them or they have not considered it. This is the state of *honest epistemic ignorance*: low T, high I, low F. The classical scale records it identically to ambivalence, which is its functional opposite.

State 3 — Strategic evasion. The respondent has an opinion but chooses not to express it, typically for reasons of institutional or social context: fear of professional consequences, unwillingness to conflict with colleagues, perception that the survey is not truly anonymous despite assurances. Strategic evasion produces a point 3 whose origin is not epistemic but political. It is the state that agnotology (Proctor and Schiebinger, 2008) studies under the name of manufactured ignorance.

State 4 — Genuine indifference. The respondent understands the question, has sufficient information, but finds the topic irrelevant to their practice. This is neutrality in its full sense. T and F are both low not from lack of information but from lack of affective or practical engagement with the topic.

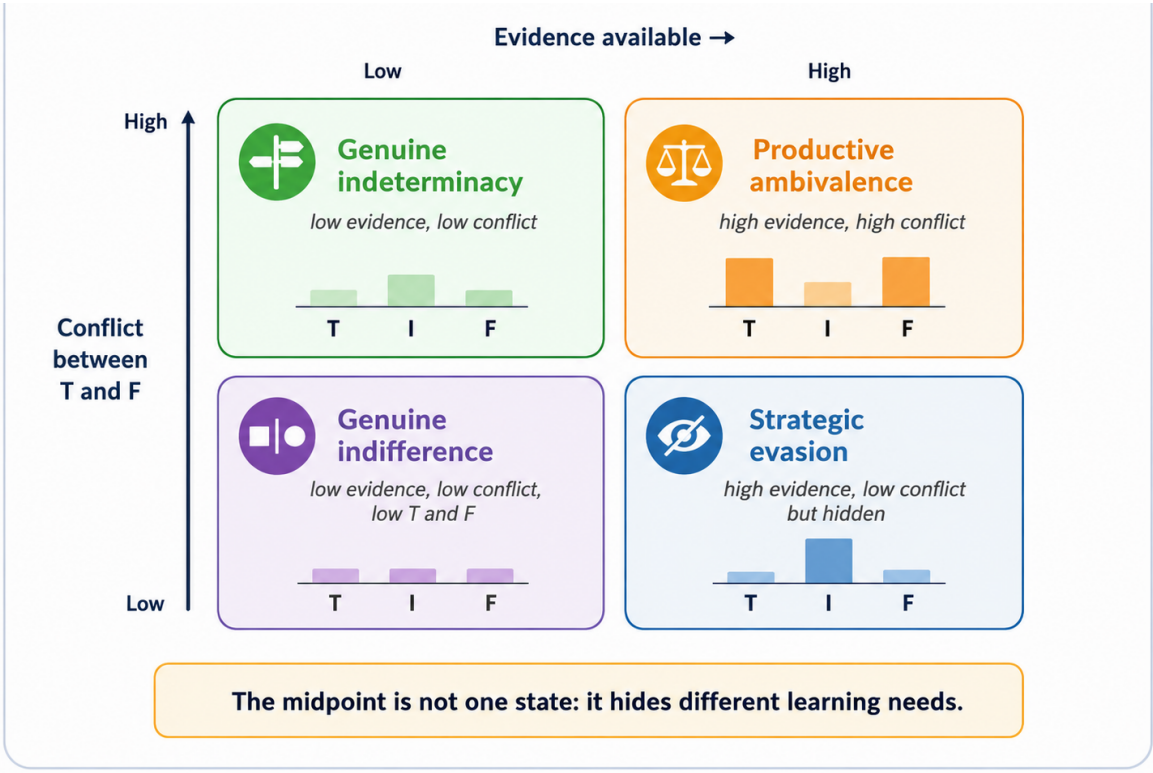

Figure 6.2. A 2×2 matrix: horizontal axis is "evidence available" (low/high); vertical axis is "conflict between T and F" (low/high). Each quadrant shows one of the four midpoint states with its TIF profile. Quadrant labels: top-left = Genuine indeterminacy (low evidence, low conflict); top-right = Productive ambivalence (high evidence, high conflict); bottom-left = Genuine indifference (low evidence, low conflict, low F and T); bottom-right = Strategic evasion (high evidence, low conflict but hidden).

Table 6.1. The four midpoint states and their neutrosophic profile.

State	T	I	F	Classical reading	Required treatment
Productive ambivalence	high	low–medium	high	Neutral (point 3)	Preserve contradiction; investigate conditions
Genuine indeterminacy	low	high	low	Neutral (point 3)	Signal information gap; provide sources

Strategic evasion	low–medium	high	low	Neutral (point 3)	Triangulate with indirect questions; agnotological analysis
Genuine indifference	low	low	low	Neutral (point 3)	Weight item; possible exclusion from analysis

6.2.2 Why the Problem Is Structural, Not a Sample Size Issue

A frequent error when diagnosing this problem is to attribute it to small samples or poorly worded questions. Neither of the two classic remedies solves midpoint collapse because the problem is not statistical: it is representational.

Increasing the scale from 5 to 7 or 10 points does not help. If the respondent is in a state of productive ambivalence (T = 0.8, F = 0.7), no point on a one-dimensional continuum correctly represents their state; they will simply move the marker to point 4 or 6 depending on which emotion predominates at the moment of responding. The scale has gained apparent granularity but not representational capacity for paraconsistent states.

Better question wording helps reduce semantic ambiguity, which is a different and legitimate problem. But a perfectly clear item can generate all four midpoint states. The problem is that the Likert scale has one dimension when attitudinal reality has at least three independent ones: the degree to which the proposition is supported by the respondent's experience (T), the degree to which it generates uncertainty or incomplete information (I), and the degree to which it is contradicted by their experience (F). Compressing three dimensions into one is not an innocent decision; it is a decision that discards information by construction.

6.3 The Solution: TIF as Decomposition of the Midpoint

The neutrosophic extension of the Likert scale replaces a one-dimensional question with a system of three independent measures that together characterise the complete attitudinal state.

The formalisation is direct. Let p be an attitudinal proposition (for example: "The use of generative AI in the classroom improves the learning outcomes of my students"). The

classical Likert scale assigns to p a value in the ordered set {1, 2, 3, 4, 5}. The neutrosophic Likert scale assigns to p a triple:

where:

- is the degree to which the respondent has evidence or experience that *supports* the proposition.
- is the degree of *indeterminacy*: missing information, ambiguity, unresolved uncertainty.
- is the degree to which the respondent has evidence or experience that *contradicts* the proposition.

The critical constraint is that **no restriction** is imposed on . The sum can take any value in . This is what distinguishes the neutrosophic formulation from its predecessors:

- If and : the triple collapses to the classical Likert scale (special case).
- If and : genuine indeterminacy (missing information).
- If : the paraconsistent case, productive ambivalence — not representable in classical Likert.

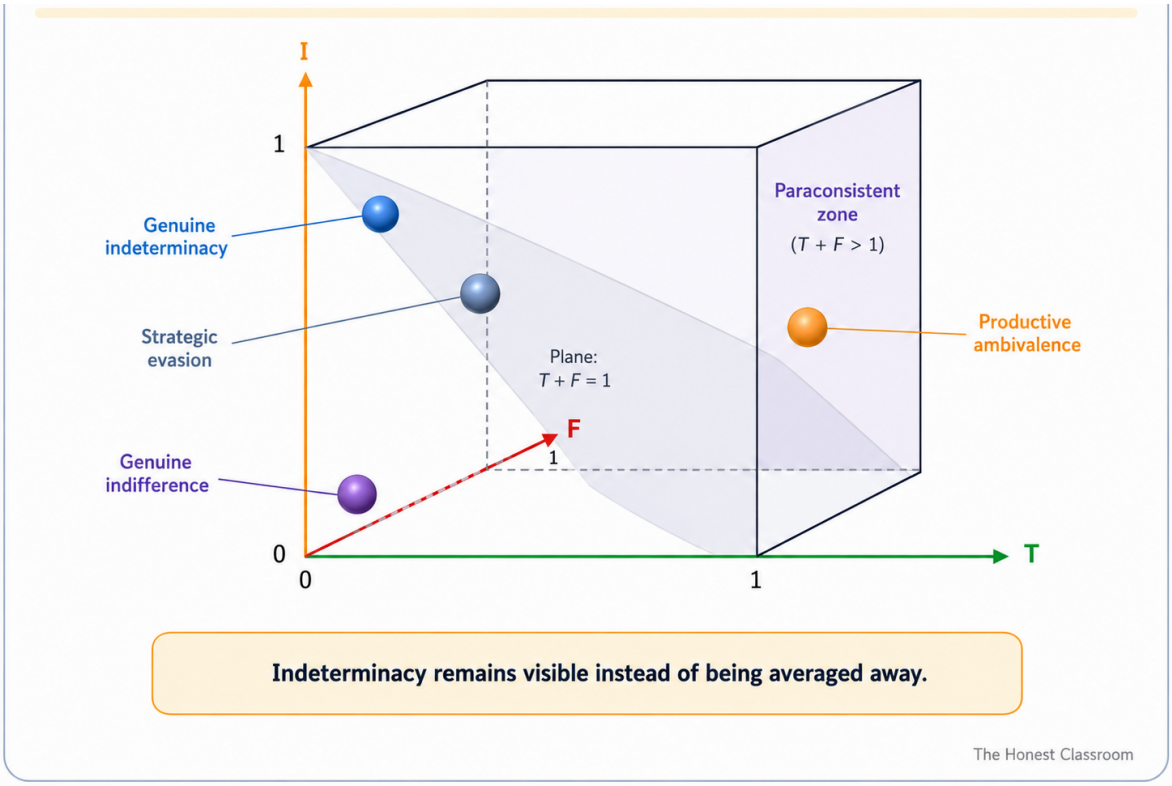

Figure 6.3. A unit cube [0,1]³ with axes T, I, F. Four coloured spheres are placed at the representative coordinates of the four midpoint states. A diagonal plane T+F=1 (semi-transparent) cuts the cube, with the region T+F>1 highlighted as the "paraconsistent zone." Labels point to each sphere.

The four states from the previous section are now expressed with precision:

State	Typical TIF profile	TIF sum	Neutrosophic zone
Productive ambivalence	(0.75, 0.20, 0.70)	1.65	Contradiction
Genuine indeterminacy	(0.15, 0.80, 0.10)	1.05	Ignorance
Strategic evasion	(0.20, 0.75, 0.15)	1.10	Ignorance (with contextual signal)
Genuine indifference	(0.10, 0.15, 0.10)	0.35	Deep ignorance

6.4 Operationalisation: How to Administer the Neutrosophic Likert Scale

There are three ways to administer the neutrosophic instrument, with different cognitive demands on the respondent and different analytical advantages.

6.4.1 Direct Format: Three Parallel Questions

The most transparent format reformulates each classical item into three explicit questions, one per dimension:

> ***Original item (classical):***
> ****The use of generative AI in the classroom improves the learning outcomes of my students.****
> ***1 2 3 4 5 Neutrosophic reformulation (direct format):***
> ***(T) To what degree does your experience or knowledge *support* this statement?***
> ***0 ——— 0.25 ——— 0.50 ——— 0.75 ——— 1.0***
> ***(I) To what degree do you have *uncertainty* or incomplete information about this statement?***
> ***0 ——— 0.25 ——— 0.50 ——— 0.75 ——— 1.0***
> ***(F) To what degree does your experience or knowledge *contradict* this statement?***
> ***0 ——— 0.25 ——— 0.50 ——— 0.75 ——— 1.0 Advantages: maximum***

representational fidelity; the respondent directly expresses all three attitudinal dimensions.

Limitations: the instrument triples in length; some respondents are confused by the possibility of marking high values for both T and F simultaneously; requires brief prior instruction.

When to use: studies that prioritise measurement depth over instrument length; respondent groups with sufficient methodological training; research where the paraconsistent case is of special interest.

6.4.2 Indirect Format: Post-hoc Conversion with LLM

The indirect format preserves the classical five-point instrument but adds an optional open question:

> **Use this space if you would like to add any clarification to your previous answer:**
> _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ *The qualitative comments are then processed by a language model with a structured prompt (included in Lab 6.A) that extracts the implicit TIF triple from the text. The numerical value of the classical scale provides the first reading; the qualitative comment provides the neutrosophic correction when the respondent chose to elaborate.*

Advantages: compatible with already-designed and validated instruments; does not increase respondent cognitive load beyond an optional box.

When to use: longitudinal studies with previously validated instruments; large-scale surveys where redesigning the instrument is not viable.

6.4.3 Hybrid Format: Classical Scale + Certainty Scale

The hybrid format adds a second question about the I dimension only:

> ***Attitude question (scale 1–5):***
> ****The use of generative AI in the classroom improves the learning outcomes of my students.****
> *□1 □2 □3 □4 □5*
> ***Certainty question (scale 1–5):***
> ****How certain are you of the previous answer?****
> ***Not at all Slightly Moderately Fairly Very certain***
> ***Conversion equations:***

Let be the attitude value and the certainty value.

A respondent who marks 3 on attitude and 1 on certainty produces (0.5, 1.0, 0.5) — a case of high ambivalence with maximum uncertainty, exactly what the classical instrument recorded as a single "neutral."

When to use: surveys to teachers or students without methodological training; studies where the main interest is distinguishing formed from provisional opinion.

6.5 Analysis Pipeline: From Raw Data to Neutrosophic Zone

Once data are collected in any of the three formats, the analysis pipeline follows four steps: parsing, conversion, zone assignment, and aggregation.

6.5.1 Step 1 — Parsing and Conversion to TIF Triple

For the direct format: the triple is directly available. Each respondent produces a vector .

For the hybrid format: conversion equations as above.

For the indirect format: the LLM processes the qualitative comment with the following prompt template (full version in Lab 6.A):

```
Analyse the following respondent comment to an attitude questionnaire
and extract the implicit neutrosophic triple (T, I, F).

T = degree to which the text expresses SUPPORT or FAVOURABLE evidence [0.0–1.0]
I = degree of uncertainty, doubt, conditionality, or missing information
[0.0–1.0]
F = degree to which the text expresses REJECTION or CONTRARY evidence [0.0–1.0]

Do NOT impose T+I+F = 1.
If T+F > 1: report PARACONSISTENT CASE.
If I > 0.6: report GENUINE INDETERMINACY.

Item context: [INSERT ITEM TEXT]
Respondent comment: [INSERT COMMENT]

Reply ONLY with this JSON: {"T": 0.XX, "I": 0.XX, "F": 0.XX,
"type": "normal|paraconsistent|indeterminacy", "justification": "one line"}
```

6.5.2 Step 2 — Neutrosophic Zone Assignment

Each triple is classified into one of four zones following the taxonomy of Chapter 1:

Zone	Classification criterion	Interpretation
Consensus	, ,	Consolidated favourable attitude with low uncertainty
Dissent	, ,	Consolidated unfavourable attitude with low uncertainty
Ambiguity	(any T, F)	High uncertainty; provisional opinion or lack of information
Contradiction	**and**	Paraconsistent state; productive ambivalence

6.5.3 Step 3 — Summary Statistics

Group mean triple:

Zone distribution: count and proportion of respondents in each zone.

Polarisation Index:

A PI close to 1 indicates high polarisation (T and F nearly equal). A PI close to 0 indicates clear predominance of one dimension. This index has no equivalent in classical Likert statistics because the classical scale cannot represent states where T and F are simultaneously high.

Certainty Index:

6.6 Human–AI Validation: Procedure and Metrics

When the LLM conversion pipeline is used (indirect or mixed format), validation follows a three-step protocol.

6.6.1 Calibration Sample

20% of comments (minimum 30 units) are randomly selected. Two human coders independently assign TIF triples without knowing the LLM's assignment.

6.6.2 Inter-rater Agreement

Agreement is measured component-by-component with Cohen's kappa adapted for continuous variables (Fleiss, Levin and Paik, 2003):

where averages observed agreement on T, I, and F separately (using ±0.25 tolerance bands to declare agreement on the continuum).

Minimum acceptable threshold for publishable research: (substantial agreement, Landis and Koch, 1977). Values between 0.41 and 0.60 indicate need to revise the prompt or coding criteria before proceeding.

6.6.3 Discrepancy Resolution

When the LLM and human coders diverge by more than 0.25 in any dimension, the case is reviewed in a joint session of the two human coders. The final triple is the human consensus, not the LLM's. A grave discrepancy rate (> 0.50 in any dimension) above 15% of the calibration sample is reported as an instrument limitation.

6.7 Case Application: Teacher Attitudes Toward AI Integration at UBE

6.7.1 Context and Sample

In the second semester of 2024, a neutrosophic Likert instrument (direct format) was administered to 30 teachers at the Universidad Bolivariana del Ecuador, distributed across four departments: Exact Sciences (n=8), Social Sciences (n=9), Humanities (n=7), and Teacher Training (n=6). Item 4 is presented here as the most methodologically informative:

> **Item 4: AI-assisted assessment allows for a fairer measurement of students' real competencies.**

6.7.2 Classical vs. Neutrosophic Results

Table 6.2. Distribution of neutrosophic zones for Item 4 (n = 30).

Zone	n	%	Group mean triple

Favourable consensus	8	26.7%	(0.78, 0.18, 0.12)
Dissent	6	20.0%	(0.11, 0.20, 0.74)
Ambiguity	9	30.0%	(0.22, 0.79, 0.18)
Contradiction (paraconsistent)	7	23.3%	(0.71, 0.30, 0.68)

Overall mean triple: (0.47, 0.38, 0.42)

Polarisation Index: 0.89 (very high)

Certainty Index: 0.62 (moderate)

Policy implication: the classical scale would have suggested that the 47% favourable majority is sufficient to proceed. The neutrosophic scale reveals that only 26.7% hold a genuinely consolidated attitude; the remainder is split between ignorance (which resolves with information) and ambivalence (which does not resolve with persuasion but with participatory design). The policy recommendation changes: before implementing AI-assisted assessment, the leadership team must distinguish between the 9 teachers who need training and the 7 who need to participate in system design.

6.8 Agnotology: When High I Is Not Honesty but Strategy

The neutrosophic triple represents the respondent's declared state, not their real state. Robert Proctor (2008) documents how industries, institutions, and political actors deliberately construct zones of uncertainty to delay decisions. In educational contexts, the mechanism operates at smaller scale but with the same structure: a teacher who fears being negatively evaluated for showing resistance will mark high I not because they do not know what they think but because they know it is not convenient to express it.

Three triangulation procedures to distinguish genuine epistemic indeterminacy from strategic evasion:

a) Paired questions: one direct question about the attitude and one indirect question framed in the third person ("What do you think your colleagues generally think about this?"). A large divergence between the direct response (high I) and the response about colleagues (which typically reveals the respondent's own opinion projected outward) signals strategic evasion.

b) Pattern analysis: if a respondent reports high I only on politically or institutionally sensitive items, and clear T or F on neutral items, the pattern signals selective strategic silence.

c) Qualitative triangulation: brief interviews with a subsample (10–15%) to contrast the declared triple with oral discourse, which is rarely as strategic as the written response.

Decision rule for the final report: if triangulation detects evidence of systematic strategic evasion in more than 20% of high-I cases, the researcher must report this finding explicitly and discuss its implications for instrument validity.

6.9 Comparison with Alternative Instruments

Semantic differential scale (Osgood, Suci and Tannenbaum, 1957): measures attitudes on pairs of opposing adjectives. Captures multiple dimensions but the dimensions are predefined and do not correspond to the TIF framework. Does not capture independence between T and F nor the paraconsistent case.

Dual confidence scales (Lichtenstein and Fischhoff, 1977): separate the direction of the attitude from its subjective intensity. The closest antecedent to the hybrid format of Section 6.4.3. Key difference: dual confidence scales assume certainty is one dimension; the neutrosophic framework decomposes it into uncertainty about T and uncertainty about F separately.

Fuzzy Likert (De Cock and Cornelis, 2002; Hesketh et al., 1988): allows fuzzy responses instead of point responses. Captures imprecision but maintains the constraint : does not allow paraconsistent states.

Conclusion: the neutrosophic Likert scale is the only alternative reviewed that (a) represents the paraconsistent case, (b) maintains T, I, and F as independent dimensions without sum restriction, and (c) is operationalisable with paper or digital instruments without special technical requirements.

Lab 6: Neutrosophic Likert Pipeline in Python

Full prompt in `_assets/prompts/likert_tif_extractor.txt`. Full Python code produces:

(1) Zone classification function, (2) hybrid-to-TIF conversion, (3) group summary statistics with PI and CI, (4) T-vs-F scatter plot with paraconsistent quadrant highlighted.

6.10 Exercises

Introductory 1. Take a classical Likert questionnaire of 10 items that you have applied or found in the literature. Reformulate 3 items in the direct neutrosophic format. Apply them to 5 colleagues and compare the TIF profiles obtained with what you would have recorded in the classical instrument. Are there paraconsistent cases?

Introductory 2. Choose an educational research article reporting Likert results. Identify the items with the highest proportion of "neutral" responses. Use the Lab 6.A prompt (LLM extraction) on any qualitative comments in the article, if present. Does the inferred TIF profile change the study's interpretation?

Intermediate 3. Design a neutrosophic Likert instrument of 8 items on a topic in your specialty. Apply it to 15–20 people using the hybrid format. Run the Lab 6 pipeline. Report: mean triple, zone distribution, number of paraconsistent cases. Discuss one practical implication not visible with the classical instrument.

Advanced 4. Design an agnotological protocol to distinguish between strategic evasion and genuine indeterminacy in your 8-item instrument. Implement paired questions (Section 6.8). Compare high-I patterns between items with institutional loading and without. What do you conclude?

6.11 For the Seminar

This chapter corresponds to **Encounter 3** of a postgraduate research seminar. Suggested duration: 3 hours.

Time	Activity
0–30 min	Case discussion: the university that decided to implement AI based on 61% "favourable." What would have changed with the neutrosophic result?
30–60 min	Small groups: design 3 neutrosophic items on their own research topic (direct or hybrid format)
60–90 min	Live demo of Lab 6 pipeline with group data
90–120 min	Agnotological discussion: in what contexts of your field is strategic evasion a serious

	methodological problem?
120–180 min	Each participant writes a triangulation protocol for the most sensitive item in their instrument

References

Atanassov, K. T. (1986). Intuitionistic fuzzy sets. Fuzzy Sets and Systems, 20(1), 87–96. https://doi.org/10.1016/S0165-0114(86)80034-3

De Cock, M., Cornelis, C., & Kerre, E. E. (2007). Fuzzy rough sets: The forgotten step. IEEE Transactions on Fuzzy Systems, 15(1), 121–130. https://doi.org/10.1109/TFUZZ.2006.889762

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions (3rd ed.). Wiley.

Hesketh, B., Pryor, R., Gleitzman, M., & Hesketh, T. (1988). Practical applications and psychometric evaluation of a computerized fuzzy graphic rating scale. Advances in Psychology, 56, 425–454. https://doi.org/10.1016/S0166-4115(08)60493-8

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? Organizational Behavior and Human Performance, 20(2), 159–183. https://doi.org/10.1016/0030-5073(77)90001-0

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 22(140), 1–55.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. University of Illinois Press.

Proctor, R. N., & Schiebinger, L. (Eds.). (2008). Agnotology: The making and unmaking of ignorance. Stanford University Press.

Smarandache, F. (1998). Neutrosophy: Neutrosophic probability, set, and logic: Analytic synthesis & synthetic analysis. American Research Press.

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Wang, H., Smarandache, F., Zhang, Y. Q., & Sunderraman, R. (2010). Single valued neutrosophic sets. Multispace and Multistructure, 4, 410–413.

Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Chapter 7. What the Text Does Not Say

Neutrosophic Sentiment and Stance: measuring ambivalence and silence

"Language is the instrument with which human beings conceal themselves as much as they reveal."

— Inspired by Robin Lakoff, Language and Woman's Place (1975)

7.1 Two Conversations About the Same Policy

In October 2024, the administration of a secondary school in Guayaquil notified parents of the implementation of an AI tutoring system starting the following semester. The communication generated an online forum of 147 messages over four days.

A research team analysed the messages with a standard sentiment analysis tool: 54% were classified as "positive," 18% as "negative," and 28% as "neutral." The administration interpreted the result as majority support and proceeded without further consultation.

Six weeks later, 43% of parents had submitted written requests to exclude their children from the system. The organised protest specifically mentioned that "nobody had really asked them."

The sentiment analysis had not lied. It had correctly classified the surface emotional polarity of the texts. What it had not done was distinguish between a message that said "great that the school is innovating" (genuine positive sentiment) and one that said "I suppose something has to be done with technology" (resignation coded as positive by the slightly affirmative tone). Nor had it detected the sixteen messages that took no explicit position — marked "neutral" — but which, read in context, were strategic silences from parents who distrusted the tool but did not want to appear "against progress" in a public forum named by the institution.

This chapter extends the TIF framework to text analysis: first to sentiment (§7.2–5.5) and then to stance (§7.6–5.9). The central argument is the same as in the previous chapter, applied now to language rather than to scales: the absence of explicit polarity is not a simple datum but an aggregate of states that require decomposition to be correctly interpreted.

7.2 Classical Sentiment Analysis and Its Limits

Sentiment analysis assigns affective orientation to text (Liu, 2012; Pang and Lee, 2008). The standard formulation assigns each fragment one of three categories: positive, negative, neutral. More sophisticated models work with a continuum in .

Its limits are known but frequently ignored in educational research:

Limit 1 — Neutrality collapses distinct states. A text with score 0.0 can be: (a) genuinely neutral — without affective charge — (b) ambivalent — with positive and negative charge of equal intensity cancelling out — or (c) ironic — with apparent polarity opposite to the real one. The standard model records all three identically.

Limit 2 — Irony and sarcasm invert the signal. "Sure, because giving my ten-year-old a robot to learn from is exactly what they needed" has a positive score in most classical models due to the accumulation of words with positive valence. Ironic inversion requires pragmatic understanding that lexical polarity models do not have.

Limit 3 — Mixed emotions have no representation. Nostalgia, affective ambivalence, enthusiasm with reservations: all are states where positive and negative valences coexist. A parent who writes "I'm excited that my daughter has these tools although I'm worried I don't know how to supervise them" has simultaneously high positive and high negative charge. The classical score collapses it into an intermediate value that erases the real affective structure.

7.3 Neutrosophic Sentiment: Ambivalence as Data

The neutrosophic extension replaces the one-dimensional score with an independent triple:

where for a document in relation to an attitudinal object :

- is the degree to which expresses substantively positive sentiment toward .
- is the degree of *affective indeterminacy* (ambiguity, conditionality, unresolved reservation, expressed uncertainty, unresolved irony).
- is the degree to which expresses substantively negative sentiment toward .

7.3.1 The Cases Where Neutrosophic Sentiment Differs from Classical

Table 7.1. Cases where neutrosophic sentiment differs from classical analysis.

Case	Classical score	TIF triple	Zone	Policy implication

Productive ambivalence	~0.2 (neutral-positive)	(0.75, 0.20, 0.70)	Contradiction	Participatory design; not persuasion
Irony	~0.4 (positive)	(0.05, 0.20, 0.90)	Dissent	Signal of concealed rejection; high risk
Conditional enthusiasm	~0.6 (positive)	(0.60, 0.55, 0.10)	Ambiguity	Conditional support; map conditions
Genuinely neutral text	0.0 (neutral)	(0.05, 0.10, 0.05)	Deep ignorance	No attitudinal information; exclude from analysis

7.4 Neutrosophic Sentiment Pipeline

The pipeline combines a standard NLP model for the base polarity signal with an LLM for complete neutrosophic annotation. The NLP model is fast and scalable; the LLM captures the I dimension and detects cases where the classical signal is inverted.

7.4.1 Step 1 — Pre-processing and Segmentation

Text is segmented into analysis units. For educational forums, the recommended unit is the complete message (not the sentence), because ambivalence is frequently constructed across multiple sentences.

7.4.2 Step 2 — Base Polarity Signal (NLP Model)

For Spanish texts: `pysentimiento` (Pérez et al., 2021) or `cardiffnlp/twitter-xlm-roberta-base-sentiment`. The score serves as a first approximation and as context for the next step.

7.4.3 Step 3 — Neutrosophic Annotation with LLM

For texts where the base polarity score is high (|score| > 0.7), the LLM confirms or corrects; for texts with score near 0 (|score| < 0.3), the LLM determines whether the case is productive ambivalence, genuine indeterminacy, irony, or text without opinion.

7.4.4 Step 4 — Cross-validation

Same protocol as Chapter 6: 20% sample, two human coders, neutrosophic kappa. Acceptance threshold: $\kappa \geq 0.61$. For irony specifically, recommend increasing calibration sample to 30%.

7.5 Case Application: Parent Forum on AI Education Policy (n=147)

Table 7.2. Classical vs. neutrosophic comparison — parent forum (n=147).

Analysis	Result	Usual interpretation
Classical	54% positive / 18% negative / 28% neutral	Majority support — proceed
Neutrosophic	(see Table 7.3)	See disaggregated interpretation

Table 7.3. Distribution of neutrosophic zones — parent forum (n=147).

Zone	n	%	Mean triple	Description
Genuine consensus	31	21.1%	(0.80, 0.15, 0.08)	Consolidated genuine support
Dissent	22	15.0%	(0.08, 0.18, 0.82)	Consolidated rejection (includes ironic messages misclassified by classical model)
Ambiguity	48	32.7%	(0.28, 0.76, 0.22)	Conditional support or uninformed
Contradiction	29	19.7%	(0.72, 0.25, 0.68)	Productive ambivalence
No opinion	17	12.6%	(0.06, 0.09, 0.07)	Logistical messages without attitudinal

				content

Key conclusion: the classical model correctly counted 39% of its "positives" as genuine support. The remaining 61% represented attitudinal states requiring different policy responses.

7.6 Neutrosophic Stance: Silence as Political Data

7.6.1 Classical Stance

Stance asks "is the author of this text in favour or against X?" rather than "what emotion does this text convey?" The standard formulation classifies each text in one of three categories regarding a target proposition: FAVOUR, AGAINST, NONE.

7.6.2 The Three Inhabitants of NONE

State 1 — Epistemic abstention: the author genuinely lacks sufficient information. Their NONE is honest and temporary.

State 2 — Genuine neutrality: the author knows the proposition and genuinely has no position — they are indifferent, or consider both options equally valid. Their NONE is stable and does not resolve with persuasion.

State 3 — Strategic silence: the author has a formed position but chooses not to express it because the context makes expression costly. Their NONE conceals a real FAVOUR or AGAINST.

Strategic silence is the most politically informative state and the most methodologically invisible. It is the state that explains why 28 of the 147 parents who wrote "neutral" in the forum — 19% — organised the protest six weeks later.

7.6.3 Neutrosophic Stance: Extended NLI

where:

- : degree to which the text supports the proposition.
- : degree to which the text contradicts the proposition.
- : degree of non-commitment regarding the proposition.

The critical difference from classical NLI is that the three values are not normalised to 1 in the final analysis. A standard NLI model imposes ; neutrosophic analysis postpones that normalisation and allows and to be simultaneously high if the text constructs a contradictory argument.

7.6.4 Identifying Strategic Silence

The neutrosophic triple alone cannot distinguish epistemic abstention from strategic silence. Three contextual signals help:

Table 7.4. Diagnostic indicators of the type of NONE.

Indicator	Epistemic abstention	Genuine neutrality	Strategic silence
Selectivity pattern	No (high I across multiple topics)	No (high I across multiple topics)	Yes (high I selective on sensitive topics)
Informative questions in text	Frequent	Absent	Absent or evasive
Declared position in post-survey interview	Formed (post-information)	Still neutral	Different from written text
Participation in subsequent collective action	Low	Low	High

7.7 Neutrosophic Stance Pipeline

The pipeline follows the same four-step logic as sentiment (§7.4), with two differences:

Difference 1: requires explicitly defining the target proposition before analysis.

Difference 2: the NLI model directly produces a signal in three dimensions (entailment, neutral, contradiction) that maps directly to (T, I, F).

Full Python code in Lab 7.B uses the `zero-shot-classification` transformer pipeline. Full stance extraction prompt in `_assets/prompts/stance_tif.txt`.

7.8 Integrated Case: Sentiment and Stance Combined in the Same Corpus

Table 7.5. Combined sentiment + stance analysis — parent forum sample (n=40, selected messages).

ID	Fragment (edited)	Sentiment TIF	Stance TIF	Notable discrepancy
M07	"Great news, finally something modern"	(0.85, 0.10, 0.05)	(0.80, 0.15, 0.05)	None — coherent
M19	"Great news..." (ironic context)	classical: (0.80, 0.20, 0.05) corrected: (0.05, 0.20, 0.85)	(0.05, 0.25, 0.85)	Yes — NLI detects F high that classical misses
M34	"Interesting if teachers control it"	(0.65, 0.50, 0.10)	(0.55, 0.65, 0.10)	Yes — I stance > I sentiment: condition is stronger in stance than in emotion
M41	"I need to talk to my husband before commenting"	(0.10, 0.85, 0.05)	(0.10, 0.90, 0.05)	None — coherent high I
M88	"The school has always made good decisions"	(0.70, 0.15, 0.10)	(0.60, 0.30, 0.40)	Yes — contradictory stance (F=0.40) invisible in sentiment

Key observation for message M88: the text has genuine positive sentiment toward the school, but stance toward the specific proposition about AI is contradictory (T=0.60,

F=0.40). The parent trusts the institution and distrusts the technology simultaneously. A sentiment-only analysis would have marked "positive" and missed the institutional/technology tension.

7.9 Cross-validation: When the Two Analyses Should Agree

Divergence Type 1 — Positive sentiment, against stance: occurs when the author expresses general enthusiasm for technology but opposes the specific implementation. The rejection is contextual, not ideological.

Divergence Type 2 — Negative sentiment, in favour stance: occurs frequently in institutional resignation. "Although I'm not fully convinced, I understand this is coming and it's better to adopt it now." The affective tone is negative; the stance is pragmatic acceptance.

Divergence Type 3 — High I in sentiment, high F in stance: the author uses careful emotional language but takes a clear position against. Typical in formal institutional communication.

Operational rule: when divergence between sentiment and stance exceeds 0.4 in the same dimension for more than 20% of the corpus, the researcher must report the divergence as a finding in itself.

7.10 Exercises

Introductory 1. Take 10 messages from an educational forum in your context. Apply the sentiment prompt (Lab 7.A) manually to 5 of them. Are there cases where classical sentiment would have produced a different reading than the TIF triple?

Introductory 2. Choose a news article about AI in education in a regional newspaper. Define the target proposition. Apply the stance prompt (Lab 7.C) to 5 reader comments. Classify each into its neutrosophic zone and identify if there are strategic silences.

Intermediate 3. Collect 30 messages from a forum or survey on an educationally relevant topic in your institution. Run the Lab 7 pipeline. Compare the zone distribution of sentiment with that of stance. Are there systematic divergences? What do they indicate?

Advanced 4. Design a protocol to detect strategic silence in your corpus. Implement at least two of the three signals from Section 7.6.4. Apply it to the corpus from Exercise 3. How many of the high-I cases are probably strategic silence vs. genuine epistemic abstention?

References

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 632–642). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1075

Küçük, D., & Can, F. (2020). Stance detection: A survey. ACM Computing Surveys, 53(1), Article 12, 1–37. https://doi.org/10.1145/3369026

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167. Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). Stance and sentiment in tweets. ACM Transactions on Internet Technology, 17(3), Article 26, 1–23. https://doi.org/10.1145/3003433

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alonso Alemany, L., & Martínez, M. V. (2021). pysentimiento: A Python toolkit for opinion mining and social NLP tasks. arXiv. https://arxiv.org/abs/2106.09462

Proctor, R. N., & Schiebinger, L. (Eds.). (2008). Agnotology: The making and unmaking of ignorance. Stanford University Press.

Smarandache, F. (1998). Neutrosophy: Neutrosophic probability, set, and logic: Analytic synthesis & synthetic analysis. American Research Press.

Wang, H., Smarandache, F., Zhang, Y. Q., & Sunderraman, R. (2010). Single valued neutrosophic sets. Multispace and Multistructure, 4, 410–413.

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1: Long Papers (pp. 1112–1122). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Chapter 8. Whose Words Are These?

Academic integrity, AI disclosure, and the epistemic contract with the reader

"AI tools cannot meet the requirements for authorship as they cannot take responsibility for the submitted work."

— Committee on Publication Ethics (COPE), Authorship and AI tools

8.1 The Question Nobody Has Answered Clearly

Ask any group of university faculty the following three questions and count how many give the same answer:

14. Is it acceptable to use an AI to clean up the grammar of your research paper?
15. Is it acceptable to use an AI to generate the first draft of a literature review section?
16. Is it acceptable to list an AI as a co-author?

In every workshop where the author has posed these questions, the responses split roughly three ways: yes to all three, yes to one and no to two, or it depends. Nobody agrees on where the line is. Most people agree there is a line somewhere.

That absence of consensus is not moral failure — it is the predictable result of a genuinely new situation that existing ethical frameworks were not designed to handle. The APA Publication Manual (7th edition, 2020) was finalised two years before ChatGPT's public release. The Chicago Manual of Style's guidance on AI sources is one paragraph. The norms are being written in real time, inconsistently, by institutions that are themselves uncertain about what they want to prohibit and why.

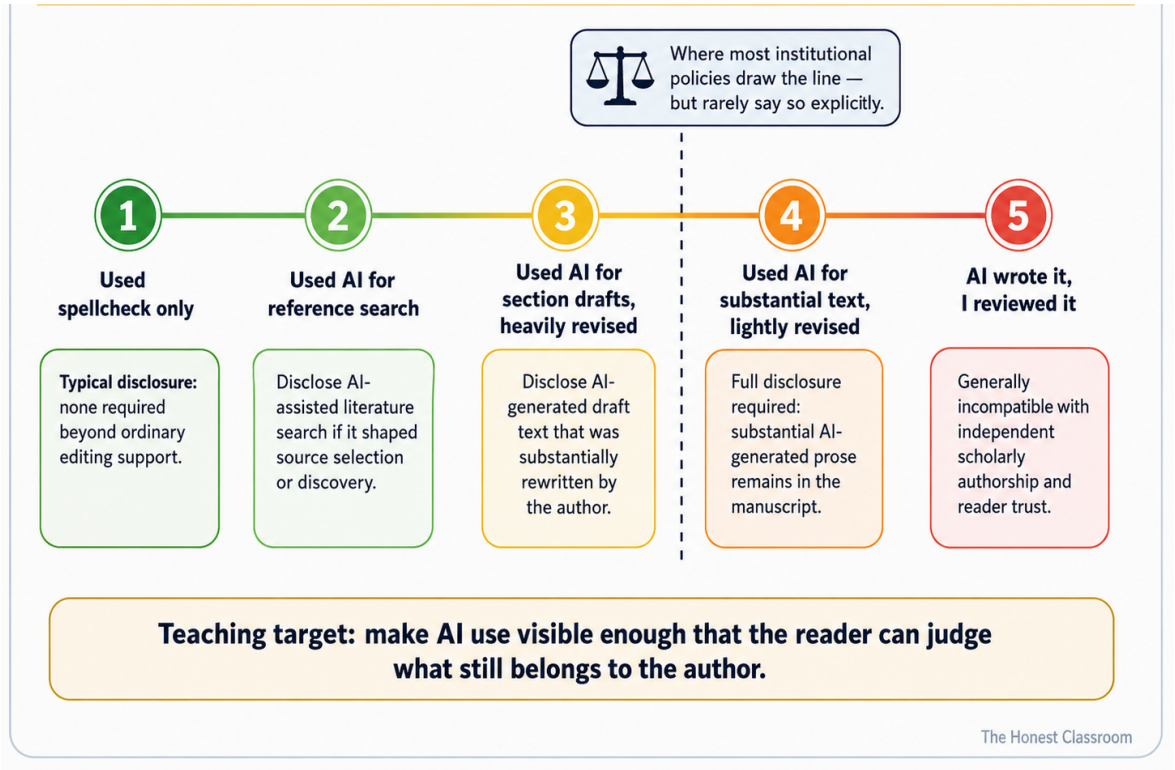
Where most institutional policies draw the line — but rarely say so explicitly.
1
2
3
4
5
Used spellcheck only
Used AI for reference search
Used AI for section drafts, heavily revised
Used AI for substantial text, lightly revised
AI wrote it, I reviewed it
Typical disclosure: none required beyond ordinary editing support.
Disclose AI-assisted literature search if it shaped source selection or discovery.
Disclose AI-generated draft text that was substantially rewritten by the author.
Full disclosure required: substantial AI-generated prose remains in the manuscript.
Generally incompatible with independent scholarly authorship and reader trust.
Teaching target: make AI use visible enough that the reader can judge what still belongs to the author.
The Honest Classroom

Figure 8.1. A spectrum showing five positions on AI use in academic writing, from left to right: (1) "Used spellcheck only" (2) "Used AI for reference search" (3) "Used AI for section drafts, heavily revised" (4) "Used AI for substantial text, lightly revised" (5) "AI wrote it, I reviewed it." Each position has a disclosure statement below it. A vertical dashed line between positions 3 and 4 is labeled "Where most institutional policies draw the line — but rarely say so explicitly." The colour shifts from green (1-3) to amber (3-4) to red (4-5).

This chapter does not resolve the debate — that would require philosophical work beyond its scope. What it does is give researchers and teachers a practical framework for making disclosure decisions that are honest, defensible, and consistent with the epistemic contract between a writer and their reader.

8.2 Why the Existing Frameworks Are Insufficient

8.2.1 The plagiarism frame does not fit

Classical academic plagiarism is the presentation of another person's intellectual work as your own without attribution. The key elements are: another *person*, *intellectual work*, and *without attribution*.

AI-generated text fails on the first two. An AI model is not a person and has no intellectual rights to claim. The text it generates is not its intellectual work in any legally or philosophically robust sense — it is a statistical recombination of patterns in training data that belongs to no one and everyone simultaneously.

This means that using AI text without attribution is not plagiarism in the classical sense. It is something else: a form of epistemic misrepresentation — misleading the reader about the origin and reliability of the argument they are reading.

The distinction matters because the appropriate response to plagiarism (attribution) is different from the appropriate response to epistemic misrepresentation (disclosure of process and verification). A student who copies a paragraph from a book and does not cite it should cite the book. A student who uses an AI to generate a paragraph and presents it as their own reasoning should not cite the AI — they should either revise the paragraph until it genuinely represents their reasoning, or declare that the paragraph is AI-assisted and specify what verification they applied to it.

8.2.2 The tool frame is incomplete

The counter-argument is that AI is just a tool, like a calculator or a search engine. Nobody requires a researcher to declare "I used Google Scholar to find these references." Why should they declare "I used Claude to generate a first draft"?

The tool frame is partially correct and partially misleading.

It is correct that AI is a tool in the same sense that a calculator is a tool: it performs operations that a human could perform, faster and at lower cost. A calculator does arithmetic. An AI generates text.

It is misleading because calculators do not produce *arguments*. They produce numbers that the researcher uses in arguments. An AI, by contrast, produces text that looks like an argument — complete with claims, evidence, transitions, and conclusions. The reader who reads an AI-generated argument is reading something that presents itself as the product of a reasoning process that may not have occurred. The tool frame ignores the difference between *producing a calculation* and *producing a claim*.

The honest middle position: AI is a tool when it is used for operations that do not generate claims (grammar checking, formatting, reference formatting, translation of your own ideas). It is something more than a tool when it generates claims that appear in the work as if they were the author's reasoned conclusions.

8.3 The TIF Test for Your Own Work

The framework developed in the preceding chapters provides a surprisingly direct test for academic integrity: apply TIF to your own text.

For each substantive claim in your paper, article, or thesis, ask:

T: Do I have evidence that this claim is correct? Have I verified it against primary sources? Can I defend it in a seminar without referring to the AI output?

I: Is there genuine uncertainty here that I have not disclosed? Is the claim based on AI output that I have not independently verified? Are there alternative interpretations I have not checked?

F: Is there evidence that contradicts this claim that I have not addressed? Did the AI produce this claim confidently but without the sources I would need to trust it?

A claim that passes all three questions — T high, I appropriately acknowledged, F considered — is a claim the author can defend. A claim that fails on I (appears certain but is based on unverified AI output) or on F (the author has not checked whether the claim has been refuted) is a claim that needs work before it appears in a published document.

This is not a new obligation — it is the basic obligation of any researcher. What AI does is make it easier to generate claims without having performed the reasoning that would normally create them. The TIF self-audit makes that gap visible.

8.4 The Disclosure Continuum

Not all AI use requires the same disclosure. The five-position continuum of Figure 8.1 maps to five disclosure statements that are honest, brief, and appropriate at each level.

Position 1 — Standard language tools only No AI disclosure required. Grammar checking, spellchecking, and formatting tools do not generate claims and do not require declaration beyond what is standard for any written document.

Position 2 — AI-assisted reference search Disclosure: «Reference searches were conducted using [tool], with all cited sources verified in their primary publication.» The key element: «verified in their primary publication.» A reference generated by an AI that the author has not read is not a reference — it is a guess. If the author has not read it, it should not be cited.

Position 3 — AI-assisted drafting, substantially revised Disclosure: «Sections of this paper were drafted with AI assistance (Claude/ChatGPT/[tool]) and subsequently revised to accurately represent the author's analysis. All factual claims were verified against primary sources.» The key element: «subsequently revised to accurately represent the author's analysis.» If the revision was not substantial, this statement is false. The test: can the author rewrite the section from scratch without the AI output, producing an

equivalent argument? If yes, the AI was a drafting tool. If no, the revision was not substantial.

Position 4 — AI-generated text, lightly revised Disclosure: «Sections [specify] were generated by [tool] and reviewed for accuracy and consistency. The following claims were independently verified: [list]. The following claims could not be independently verified and are marked with uncertainty flags: [list].» This is a high-disclosure standard that most journals will not yet require but that is epistemically honest. It is appropriate for preprints and working papers where the author wants to be transparent about the epistemic status of each claim.

Position 5 — AI wrote it, the author reviewed it Disclosure: «This document was generated by [tool]. The author reviewed the output for accuracy and accepts responsibility for its content. The following changes were made from the original AI output: [list].» If this disclosure cannot be completed — if the author cannot list what they changed and why — then the document should not be submitted for academic purposes. Submitting AI output that the author cannot substantively account for violates the epistemic contract between author and reader regardless of whether institutional rules explicitly prohibit it.

8.5 How to Cite AI-Generated Content (APA 7 + Chicago 17)

APA 7th Edition (2020, updated 2023 guidance):

For a conversational AI output (ChatGPT, Claude, Gemini):

```
OpenAI. (2024). ChatGPT [Large language model]. https://chat.openai.com
```

In-text: (OpenAI, 2024)

For a specific generated response that you quote:

```
OpenAI. (2024, March 15). [Response to the prompt: "Explain neutrosophic
logic in three sentences"]. ChatGPT. https://chat.openai.com
```

Important limitation: AI outputs are not reproducible — a different user at a different time will not receive the same response to the same prompt. APA's guidance therefore recommends: if you quote an AI output directly, include the full prompt and the full

response in an appendix. If you cannot reproduce the output for a reader who wants to verify it, the citation is incomplete.

Chicago 17th Edition:

For footnote/endnote style:

```
OpenAI, ChatGPT, response to "Explain neutrosophic logic," March 15, 2024,
https://chat.openai.com.
```

For bibliography:

```
OpenAI. ChatGPT. Accessed March 15, 2024. https://chat.openai.com.
```

What both styles do not solve: neither APA nor Chicago has resolved how to cite an AI that contributed to the *argument* rather than the *text*. If the AI generated a conceptual framework that the author then developed, the AI is not a source — it is an intellectual collaborator whose contribution cannot be cited in the conventional sense. The honest solution in this case is a methodological note: «The conceptual framework in Section 3 was developed in dialogue with Claude (Anthropic, 2024), revised substantially by the author, and verified against [sources].»

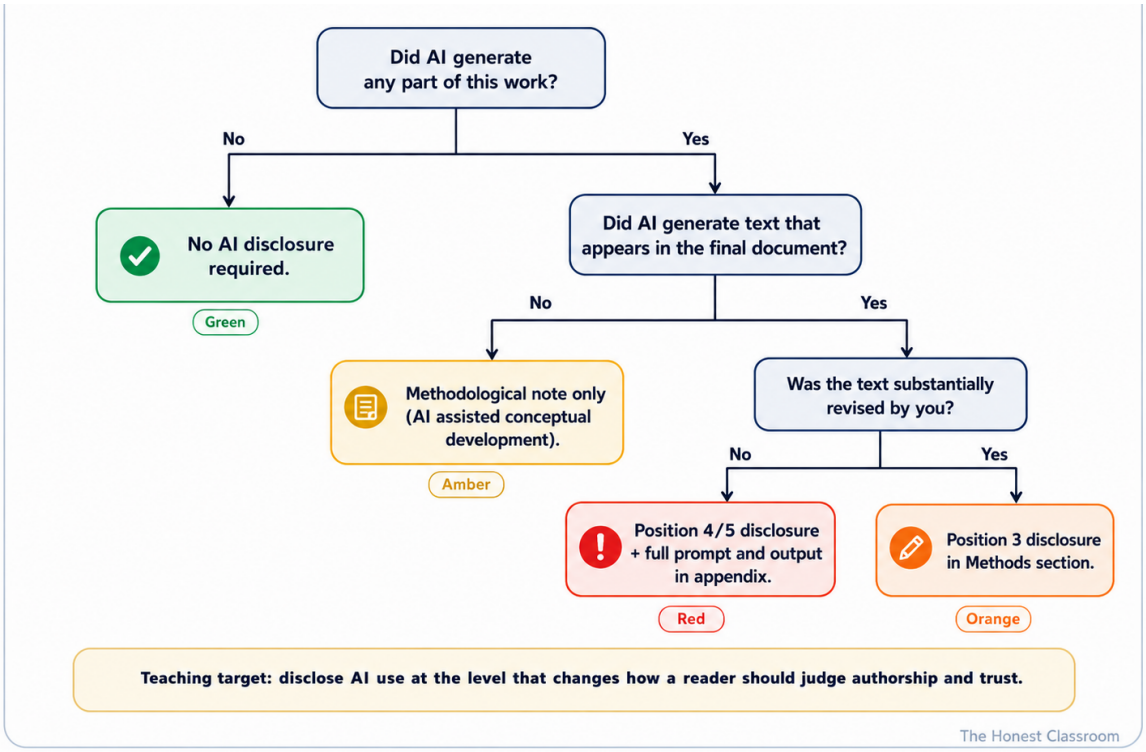
Did AI generate any part of this work?
No
Yes
No AI disclosure required.
Green
Did AI generate text that appears in the final document?
No
Yes
Methodological note only (AI assisted conceptual development).
Amber
Was the text substantially revised by you?
No
Yes
Position 4/5 disclosure + full prompt and output in appendix.
Red
Position 3 disclosure in Methods section.
Orange
Teaching target: disclose AI use at the level that changes how a reader should judge authorship and trust.
The Honest Classroom

Figure 8.2. A decision tree for AI citation decisions. Start: "Did AI generate any part of this work?" No "No AI disclosure required." Yes "Did AI generate text that appears in the final document?" No "Methodological note only (AI assisted conceptual development)." Yes "Was the text substantially

revised by you?" No "Position 4/5 disclosure + full prompt and output in appendix." Yes "Position 3 disclosure in Methods section." Each path has a coloured label (green for no-disclosure, amber for methodological note, orange for position 3, red for positions 4/5).

8.6 The Institutional Policy Gap and What You Should Do About It

Most universities currently have one of three positions on AI in academic work:

Position A — Blanket prohibition: AI is not permitted in any academic work. This position is increasingly unenforceable and creates the perverse incentive to use AI secretly rather than transparently.

Position B — Blanket permission with vague guidance: AI is permitted «as a tool» with «appropriate» acknowledgement. This position defers the ethical question to the individual author without giving them the tools to answer it.

Position C — Context-specific policy: different types of work (exams, coursework, research papers, theses) have different rules, with explicit disclosure requirements for each level. This is the most defensible position and the least common.

If your institution has Position A or B, you are not relieved of the obligation to apply the TIF test to your own work. The absence of an institutional policy does not create an ethical void — it creates individual responsibility. The standard your work should meet is: *would a reasonable colleague who knew exactly how I used AI judge my work as honest?*

If you are in a position to influence your institution's policy (as a faculty member, graduate programme director, or curriculum designer), the framework of this chapter — and specifically the five-position continuum with its disclosure statements — can serve as the basis for a concrete policy that is enforceable, pedagogically coherent, and honest about what it is trying to achieve.

8.7 Teaching Academic Integrity in the AI Era

For teachers who must explain this to students, three principles that are easier to apply than any formal policy:

Principle 1 — The seminar defence test.

If a student could not defend a claim in an oral seminar without referring to an AI output, that claim is not the student's reasoned conclusion. It can be in the paper, but it must be marked as AI-assisted and must have been verified against primary sources.

Principle 2 — The counterfactual revision test.

Ask the student: «If you deleted this section and had to rewrite it from memory, using only your own knowledge and verified sources, would you produce an equivalent argument?» If yes, the AI was a drafting tool. If no, the work needs more development before it can be claimed as the student's own.

Principle 3 — The reader's right to know.

A reader who invests time in a paper has a right to know whether the argument was reasoned by the named author or generated by a statistical model. That right does not depend on institutional policy. It is the basic epistemic respect that any author owes their reader.

References

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). American Psychological Association.

Chicago Editorial Staff. (2024). Citing content generated by artificial intelligence. The Chicago Manual of Style Online. https://www.chicagomanualofstyle.org/qanda

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. Science and Engineering Ethics, 26(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

McAdoo, T. (2023, April 7). How to cite ChatGPT. APA Style Blog. https://apastyle.apa.org/blog/how-to-cite-chatgpt

Selwyn, N. (2019). Should robots replace teachers? AI and the future of education. Polity Press.

Smarandache, F. (1998). Neutrosophy: Neutrosophic probability, set, and logic: Analytic synthesis & synthetic analysis. American Research Press.

Williamson, B., & Hogan, A. (2020). Commercialisation and privatisation in/of education in the context of COVID-19. Education International.

Chapter 9. The Grade That Does Not Fit in a Number

Neutrosophic Tensor Evaluation: measuring process, disagreement, and what the student does not yet know

“We cannot teach anyone anything; we can only help them discover something within themselves.”

— Attributed to Galileo Galilei

9.1 The Student Who Wrote Too Well

In March 2024, a law student in Bogotá submitted her final paper on comparative constitutional law. The work was brilliant: well-structured argument, verified primary sources, three pages of comparative analysis between Colombian and Spanish jurisprudence that her professor later described as "the best work of the semester." The university's AI detector flagged it at 92% probability of artificial origin.

The student was suspended pending investigation. She appealed. The investigation took three weeks and concluded she had written every word. She was simply a good writer, trained in a family of lawyers, who had learned to write with precision and without ornament since secondary school. The university reinstated her quietly. There was never a public apology. The student transferred the following semester.

The effect on the other 34 students in the cohort was informally documented by the same professor: in the two months that followed, submitted work showed measurable degradation in syntactic precision. Students had learned to write *worse* intentionally to appear more human.

This chapter is about that problem and its solution — not detection, but redesign of what is evaluated.

Figure 9.1. Opening illustration: a split image. Left half shows a student at a desk writing, with a clean essay visible; above her, a detector screen shows "92% AI-generated" in red. Right half shows a classroom where students deliberately add typos to their text, with a thought bubble showing "I must look more human." A caption: "The detector creates the behaviour it seeks to prevent."

The central argument is this: classical evaluation — based on finished products, graded with a number — was designed for a world without generative AI and does not work in the one we have. The replacement is not another detection instrument. It is a change of object: instead of evaluating *what the student produced*, evaluating *how they thought while producing it*. And to capture that *how* in all its complexity — across multiple criteria, multiple moments, and multiple perspectives — we need a mathematical object that a number cannot be. We need a tensor.

9.2 Why the Essay Died and Detectors Do Not Replace It

The academic essay has spent four centuries as a university assessment instrument. AI did not kill it because students are dishonest — the essay survived decades of paper-purchasing services. It killed it because generative AI democratised the production of plausible text to the point where the gap between "the student understands" and "the student has access to a model" is no longer visible in the final product.

Detectors are not the solution for three structural reasons:

Reason 1 — The false positive rate is unacceptable. Studies report false positive rates between 1% and 20% depending on the detector and text type (Liang et al., 2023;

Sadasivan et al., 2023). In a cohort of 200 students, a detector with a 5% false positive rate will accuse 10 innocent students per semester.

Reason 2 — The arms race is permanent and asymmetrically costly. Every improvement in detection generates improvement in evasion. The cost of the race is paid primarily by honest students who adjust their writing to avoid false accusations.

Reason 3 — They detect the symptom, not the problem. The problem is not that students use AI. It is that evaluation measures the product rather than the process, and the product can be delegated. No detector resolves that.

Table 9.1. Why classical instruments fail in the AI era.

Instrument	What it measured before	Why it fails now
Take-home essay	Written synthesis	AI writes it in 90 seconds
Multiple choice	Content recall	AI answers with 85–95% accuracy
Take-home exam	Extended reasoning	Indistinguishable from AI output with good prompting
AI detector	Presence of artificial text	1–20% false positives; trivial evasion
Oral exam	Real-time performance	Works, but does not scale and does not capture process

The last row is relevant: oral examination *does* work because it evaluates in real time, with no delegation possible. The problem is scale: it is not possible to conduct oral exams with 180 students four times per semester.

9.3 The Two-Lane Model

Safe Lane: verifies that the student can think without help. AI is prohibited. The instrument evaluates competencies that must be internalised: solving a new problem in class, orally defending a position, demonstrating a skill in real time.

Open Lane: develops professional reasoning with AI. AI is permitted and documented. The instrument evaluates whether the student can use AI with judgement: detect its errors, cite its limitations, make decisions the model cannot make.

Both lanes are necessary and non-interchangeable. A surgeon needs to know anatomy without consulting a screen (Safe Lane) and needs to know how to interpret an AI-assisted diagnosis (Open Lane).

9.4 The 4D Framework

The 4D evaluation framework organises student competence across four independent dimensions:

D1 — Delegation: did the student use AI appropriately for the task? Appropriate delegation means knowing what to delegate and what not to.

D2 — Description: did the student document their process with precision? This includes: what prompts they used, what they obtained, what they modified and why, what they discarded.

D3 — Discernment: did the student critically evaluate the AI's output? The student identifies which part of the AI's output has high T (well-founded), high I (uncertain, requires verification) or high F (incorrect or contradictory with other sources).

D4 — Diligence: did the student cite sources, verify claims, complete the task with rigour? Diligence in the AI era includes verification of AI-generated claims.

Table 9.2. 4D Rubric — performance levels.

Dimension	Excellent	Satisfactory	Insufficient	AI-dependent
D1 Delegation	Selects and justifies what to delegate	Delegates with judgement without explicit justification	Delegates randomly	Delegates everything without judgement or verification
D2 Description	Process documented, reproducible	Process partially documented	Minimal documentation	No evidence of own process
D3 Discernment	Applies TIF to output, signals errors and gaps	Identifies some errors	Accepts output with generic reservations	Accepts output without questioning

D4 Diligence	All claims verified with primary source	Majority verified	Selective verification	No verification

9.5 The Problem the 4D Framework Does Not Solve Alone

The 4D framework is a substantial improvement over classical evaluation. But it has a limitation that only becomes visible when evaluation is performed across multiple moments and by multiple people: it collapses information that should not be collapsed.

Consider Student A, who shows insufficient D3 at the start of the semester and excellent D3 at the end. Student B shows satisfactory D3 throughout. With a final rubric that averages, both might receive the same grade. But Student A learned something B did not — and that trajectory is exactly what a process-oriented evaluation system should capture.

The problem is structural: a rubric produces one number per dimension. A number cannot simultaneously represent a criterion, a time moment, and an evaluator's perspective. For that, we need an object with three modes. We need a tensor.

9.6 The Student Evaluation Tensor

A tensor is, in accessible terms, a generalisation of the matrix. A matrix has two modes: rows and columns. An order-3 tensor has three modes: like a cube where each cell is indexed by row, column, and depth.

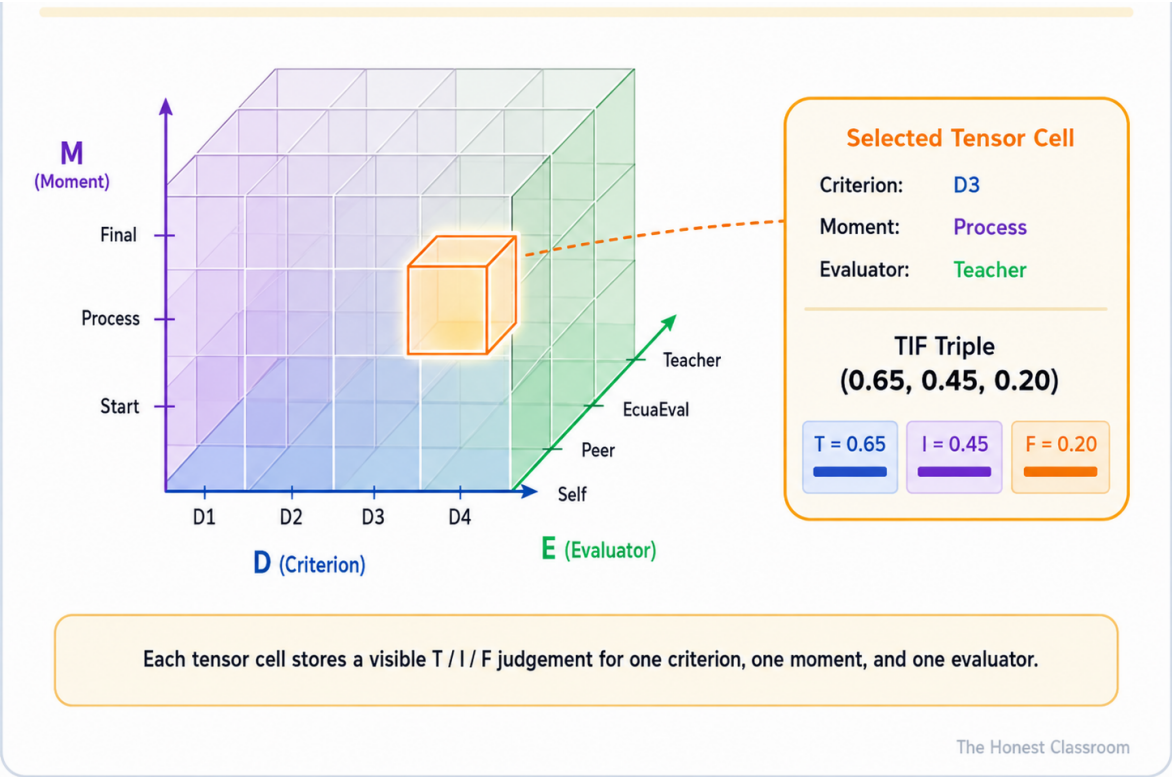

Figure 9.2. A 3D cube diagram representing the Student Evaluation Tensor (SET). Three axes labeled: D (criterion, 4 levels: D1–D4), M (moment, 3 levels: Start/Process/Final), E (evaluator, 4 levels: Self/Peer/EcuaEval/Teacher). One cell is highlighted and expanded to show a TIF triple (0.65, 0.45, 0.20) with labels for T, I, and F. The overall cube is semi-transparent to show interior structure.

The **Student Evaluation Tensor (SET)** is an order-3 tensor with neutrosophic entries:

where:

- **Mode D** (criterion): four 4D dimensions — .
- **Mode M** (moment): at least : start, process, final submission.
- **Mode E** (evaluator): : student self-assessment, peer assessment, EcuaEval (AI), teacher.

Each tensor entry is a neutrosophic triple, not a number:

9.6.1 Reading One Tensor Entry

Entry : criterion D3 (Discernment), moment M2 (process), evaluator E3 (EcuaEval).

A triple says:

- **T = 0.65:** the AI observes moderately solid evidence of critical thinking — the student signalled at least two model errors and documented them.
- **I = 0.45:** substantial uncertainty in this evaluation — the portfolio describes the errors but does not explain the reasoning for why they are errors; the AI cannot determine whether the student understood them or copied them from another source.
- **F = 0.20:** limited evidence of lack of discernment — in one passage the student accepted a model claim without verifying it.

This entry alone says more than a grade. It says: *there is real progress (T=0.65), but there is a documentation gap that prevents confirming deep understanding (I=0.45), and there is a specific case of uncritical acceptance that deserves targeted feedback (F=0.20).*

9.7 Aggregation: From Tensor to Actionable Feedback

9.7.1 Projection by Criterion: The Student's 4D Profile

The projection over mode D aggregates all moments and evaluators for each criterion:

Result: one triple per criterion — the student's **neutrosophic 4D profile**. Instead of "D3 = 7/10," the profile says "D3 = (0.62, 0.38, 0.18)": moderate performance, substantial unresolved uncertainty, isolated errors.

9.7.2 Temporal Projection: The Process Trajectory

The projection over mode M aggregates criteria and evaluators for each moment:

If grows from moment to moment and decreases, the student is consolidating their learning — exactly the pattern that Bloom and Carroll's mastery learning seeks to produce.

9.7.3 The Uncertainty Management Coefficient (UMC)

The UMC is the process summary metric, built on the tensor's I component:

where is the fall in indeterminacy between the first and last moment, and is a trajectory weight parameter with default value 0.30.

The two terms capture distinct aspects of uncertainty management:

- — *level achieved*: how much indeterminacy was resolved by the end of the period.

- — *trajectory bonus*: credit for the learning journey itself, not just the destination.

Why ? This value is not arbitrary: it corresponds to the ratio of the recommended moment weights , making the trajectory term formally consistent with the moment-weighting scheme of Section 9.7.4. Empirical calibration against independent outcome measures is needed; the parameter is adjustable in the companion library (`src/tensor_eval.py`, argument `lambda_delta`).

Level	UMC	Interpretation
Basic	0.30–0.55	Student recognises uncertainty but does not actively manage it
Intermediate	0.56–0.75	Student analyses uncertainty and documents it in their process
Advanced	0.76–1.00	Student actively reduces their I during the process and can explain how

9.7.4 Aggregation by Tensor Contraction

For a principled summary that preserves discriminative power without destroying information through averaging:

Recommended weights: D3 Discernment receives 0.35 (the differentiating competency of the AI era); D1, D2, D4 receive 0.25, 0.20, 0.20. Final moment receives 0.50; process 0.35; start 0.15. Teacher vote 0.40; EcuaEval 0.30; self-assessment 0.20; peer 0.10.

This weighted-contraction operator specialises, when the tensor reduces to a single mode, to the simplified-neutrosophic weighted averaging and geometric operators of Ye (2014); the book uses it in operative form, and the reader can check the single-mode reduction directly against Ye's published definitions.

9.8 EcuaEval: The Tensor Generator

EcuaEval (full prompt in _assets/prompts/ecuaeval_tensor.txt) generates the entries of the SET — the AI evaluator layer — for each criterion and moment.

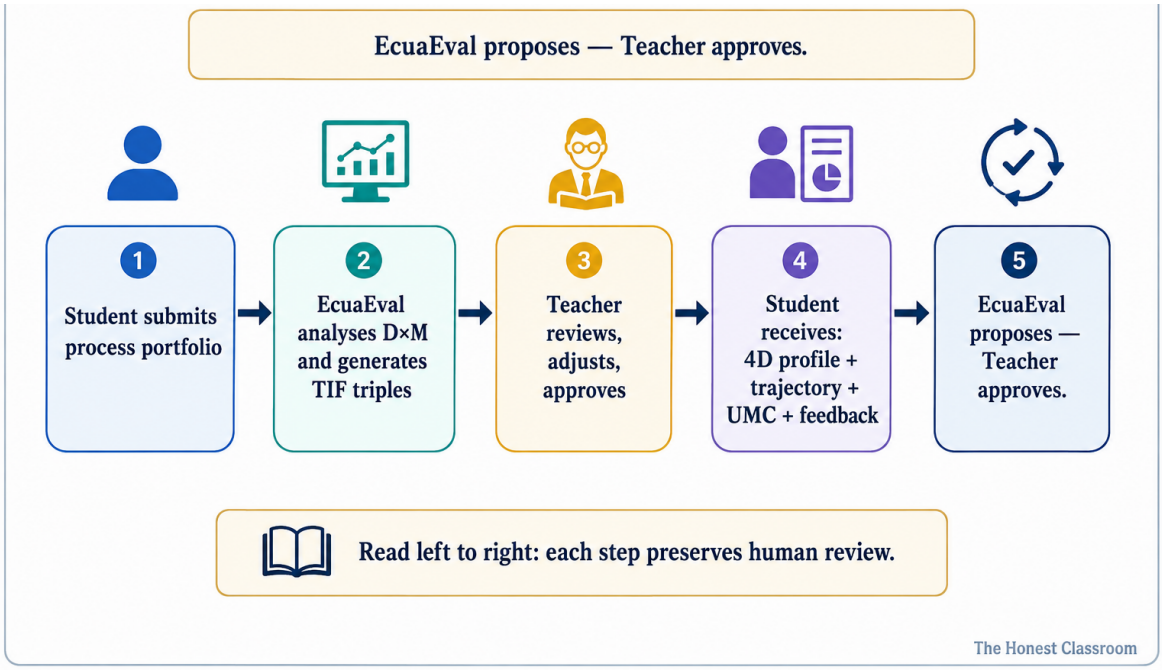
EcuaEval proposes — Teacher approves.
1
Student submits process portfolio
2
EcuaEval analyses D×M and generates TIF triples
3
Teacher reviews, adjusts, approves
4
Student receives: 4D profile + trajectory + UMC + feedback
5
EcuaEval proposes — Teacher approves.
Read left to right: each step preserves human review.
The Honest Classroom

Figure 9.3. EcuaEval process flow. The student submits a process portfolio; EcuaEval analyses the D×M structure and generates TIF triples; the teacher reviews, adjusts, and approves; the student receives a 4D profile, trajectory, UMC, and feedback. The invariant rule is: EcuaEval proposes — Teacher approves.

Invariant rule: EcuaEval generates, the teacher approves. The AI does not grade; it produces a graded proposal that human judgement confirms or corrects.

9.9 Eight Assessment Formats That Work in the AI Era

Table 9.3. Assessment formats mapped to lanes and tensor modes.

Format	Lane	Modes activated	How it works
Process portfolio	Open	Full D×M×E	Student documents prompts, outputs, decisions, reflections at each stage
Oral defence with TIF	Safe	D×M_final×E_teacher	Student presents; evaluator asks "what T, I, F do you assign

			to this claim?"
Live problem-solving	Safe	D1,D3,D4×M_single×E_teacher	Novel problem solved in class without devices
Comparative analysis	Open	D3×M_process×E_multiple	Student receives two AI responses to the same problem, evaluates with TIF
Teaching-back	Safe	D3,D4×M_final×E_peer	Student teaches the concept to a peer
Output audit	Open	D3×M_process×E_multiple	Student audits an AI output for errors by type
Structured debate	Safe	D1,D3×M_single×E_peer+teacher	Student defends a position under refutation
Build-and-explain	Open	D1,D2,D3,D4×M_complete×E_multiple	Student builds an AI agent and explains every design decision

9.10 Case Application: Tensor Evaluation in a Research Methodology Course (UBE, 2024)

9.10.1 Context

18 postgraduate students at UBE used the SET during one semester to design their thesis research instrument. Evaluation was conducted at three moments (initial proposal, prototype, final version) with four evaluators.

9.10.2 Two Representative Student Profiles

Table 9.4. Tensor profiles of two students (aggregation by criterion, final moment).

Student A — Final 4D profile:

Criterion	Final triple	Zone	Interpretation
D1 Delegation	(0.78, 0.18, 0.12)	Consensus+	Knows what to delegate; justifies well
D2 Description	(0.45, 0.55, 0.20)	Ambiguity	Process partially documented; important gap
D3 Discernment	(0.82, 0.15, 0.08)	Consensus+	Excellent; applies TIF systematically
D4 Diligence	(0.70, 0.22, 0.15)	Consensus+	Verifies sources; some URLs missing

UMC: 0.71 (Intermediate-high) | Conventional grade equivalent: 8.2/10

Student B — Final 4D profile:

Criterion	Final triple	Zone	Interpretation
D1 Delegation	(0.65, 0.30, 0.25)	Transition	Reasonable delegation without explicit criterion
D2 Description	(0.80, 0.15, 0.10)	Consensus+	Excellent process documentation
D3 Discernment	(0.40, 0.55, 0.30)	Ambiguity	Identifies some errors; does not classify them
D4 Diligence	(0.72, 0.20, 0.12)	Consensus+	Systematic verification

UMC: 0.69 (Intermediate) | Conventional grade equivalent: 7.8/10 Critical observation: conventional grades are similar (8.2 vs 7.8). The tensor reveals completely distinct strength/gap structures: Student A has a specific deficit in process documentation (D2) that their excellent discernment masks in any average; Student B has the opposite problem — perfect documentation but discernment — the critical AI-era competency — is their weakest dimension.

9.10.3 Student A's Learning Trajectory in D3

Three observations: (1) learning trajectory is clear: T grows from ~0.40 to ~0.81; I falls from ~0.61 to ~0.16. (2) Evaluator agreement increased: in M1 there was appreciable discrepancy; in M3 all four converge. (3) Self-assessment in M1 was optimistic: student rated T=0.45 while EcuaEval and teacher saw T=0.35 — the student did not yet have clarity about what was expected.

9.11 Formal Foundation: Connection to SVN Tensors

For mathematically trained readers, the structure presented in this chapter is a Single-Valued Neutrosophic Tensor (SVN tensor): the single-valued neutrosophic set of Wang, Smarandache, Zhang, and Sunderraman (2010) extended to three modes (criterion × moment × evaluator).

Four formal properties of this structure are directly applicable, each connecting the SET to a result already established in the published literature:

17. **Level-cut decomposition**: the SET can be reconstructed exactly from a parameterised family of binary cut tensors — enabling simplified binary versions and recovery of the full tensor when needed.
18. **Tucker decomposition** (Kolda and Bader, 2009): identifies the dominant modes of variation across the group — pedagogically, which criterion × moment × evaluator combination explains most variance in group performance.
19. **Contraction-based aggregation: the operator of Section 9.7.4 recovers the SVNWA and SVNWG operators of Ye (2014) as special cases when the tensor reduces to a single mode.**
20. **Reduction to prior frameworks: when throughout the tensor, the SET reduces to the intuitionistic fuzzy tensor of Chen and Chen (2019). Classical rubrics are the special case with and : the SET generalises without discarding prior work.**

These properties connect the SET to established operators in the published literature (Ye, 2014; Chen and Chen, 2019), against which they can be checked directly.

9.12 Risks, Power, and What This System Should Never Be Used For

A book that proposes an AI-assisted evaluation system without discussing its risks and power implications has done half the work. This section does the other half.

The concerns raised here are not hypothetical. Neil Selwyn (2019) has documented the ways in which digital technologies in education systematically advantage already-advantaged students and concentrate data power in institutional hands. Ben Williamson and Anna Hogan (2020) have shown how algorithmic systems in education routinely make consequential decisions about students based on data those students did not consent to generate. The Student Evaluation Tensor does not automatically escape those patterns. Without explicit constraints, it could reproduce them.

9.12.1 Four things the SET must never be used for

1. Evaluating teacher performance.

The SET measures student learning states. It is not a measure of teaching quality. A teacher whose students show high I (high uncertainty) in the early weeks of a course may be doing excellent diagnostic pedagogy — deliberately surfacing what students do not know before they falsely believe they know it.

A teacher whose students show rapid T growth may have taught to the assessment.

The tensor does not distinguish these cases, and any administrator who uses SET data to evaluate teachers has inverted its purpose. Institutional policy should prohibit this use explicitly.

2. Ranking students against each other.

The SET measures each student's trajectory relative to their own starting point, not relative to other students. A student who enters a course with high I (genuine knowledge gaps) and ends it with moderate T (partial consolidation) has learned more than their final numbers suggest. Using SET data to rank students against each other destroys the interpretive logic of the tensor: it converts a process metric back into a product metric, which is precisely what the SET was designed to avoid.

3. Generating permanent academic records without explicit consent.

The SET entries are fine-grained data about a student's epistemic state at multiple moments and from multiple evaluators. That data should live in the course and be accessible to the student and teacher. It should not be transmitted to institutional data warehouses, accreditation agencies, or third-party platforms without explicit, informed consent from the student.

The principle is simple: the data was generated to help the student learn;

it should not be repurposed to judge the student elsewhere.

4. Operating without a human-in-the-loop approval step.

EcuaEval generates proposals. The teacher approves or corrects them. This is not a bureaucratic formality — it is the load-bearing ethical constraint of the entire system. An institution that removes the teacher-approval step to save time has removed the accountability mechanism. EcuaEval is not an evaluation system; it is a decision-support system that requires a human decision-maker.

9.12.2 The structural risk: surveillance creep

The most serious risk of any learning analytics system is not its intended use but its unintended ones. A tensor that is designed to help a teacher understand a student's learning trajectory can, with minimal modification, become a surveillance instrument that monitors compliance, flags "risky"

epistemic profiles, or selects students for interventions they did not request.

Three structural safeguards against surveillance creep:

Safeguard 1 — Data minimisation. The system should collect the minimum data required for its pedagogical purpose. If a teacher does not use the concept-map dashboard (§10.5), the concept-map data should not be generated.

Collect what you use; do not collect what you might use someday.

Safeguard 2 — Student access and deletion rights. Students should be able to see their own tensor data at any time and request its deletion at the end of the course. A system that the student cannot read or revoke is a surveillance system, not a learning system.

Safeguard 3 — Institutional transparency. The teacher and institution should be able to answer, in plain language, every question a student or parent might ask about what data is collected, who sees it, how long it is kept, and what

decisions it informs. If they cannot answer those questions, the system is not ready for deployment.

9.12.3 The equity question

The SET was piloted in a postgraduate research programme at a Ecuadorian university with motivated adult students who enrolled voluntarily. That context is not the same as a large undergraduate course, a secondary school, or a low-income rural school with unreliable connectivity.

Three equity considerations that the teacher and institution must address before deploying:

Access equity: students who cannot access EcuaTutor reliably (due to connectivity, device availability, or language barriers) will have systematically incomplete SET entries. An incomplete tensor is not a neutral tensor — it is a tensor that systematically undercounts the contribution of the most disadvantaged students. The institution must ensure that access to the tutoring system is equitable before the evaluation system is equitable.

Cultural equity: the 4D criteria (Delegation, Description, Discernment, Diligence) were developed in a specific cultural context. In contexts where oral tradition, collective knowledge, and relational epistemology are primary, the bias toward documented, individual, text-based evidence (D2 and D4) may systematically disadvantage students whose knowledge practices are legitimate but different. The teacher must adapt the criteria to the cultural context of their students, not the other way around.

Language equity: EcuaTutor is currently available in Spanish and English.

Students whose first language is neither — including Kichwa, Aymara, or dozens of other languages spoken across Latin America — face an additional cognitive load in every session. This load should be acknowledged and mitigated where possible (bilingual corpus, concept map labels in the student's first language).

9.12.4 What this means for how you build the system

The constraints above are not obstacles to implementation. They are the design requirements of an ethical implementation. An EcuaTutor deployment that:

- Has teacher approval at every evaluation step
- Gives students access to their own tensor data

- Has an explicit data retention policy
- Does not use tensor data for teacher evaluation
- Addresses access, cultural, and language equity …is an honest system. A deployment that skips any of these is not neutral — it is a system that has chosen to prioritise institutional convenience over student rights, and should be named as such.

The honest classroom is not the classroom with the best tools. It is the classroom where the tools are deployed with the same epistemic honesty that the tools themselves are designed to teach.

References

Anthropic. (2026, February 23). Education report: The AI Fluency Index. Anthropic. https://www.anthropic.com/research/AI-fluency-index

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. David McKay.

Carroll, J. B. (1963). A model of school learning. Teachers College Record, 64(8), 723–733. https://doi.org/10.1177/016146816306400801

Dakan, R., & Feller, J. (n.d.). The AI Fluency Framework. AI Fluency Framework. https://aifluencyframework.org/

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. SIAM Review, 51(3), 455–500. https://doi.org/10.1137/07070111X

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? arXiv. https://arxiv.org/abs/2303.11156

Wang, H., Smarandache, F., Zhang, Y. Q., & Sunderraman, R. (2010). Single valued neutrosophic sets. Multispace and Multistructure, 4, 410–413.

Ye, J. (2014). A multicriteria decision-making method using aggregation operators for simplified neutrosophic sets. Journal of Intelligent & Fuzzy Systems, 26(5), 2459–2466. https://doi.org/10.3233/IFS-130916

Chapter 10. The Tutor That Does Not Answer

EcuaTutor: Socratic guidance, real-time TIF evaluation, and tensor generation

"And if the truth about reality is always in our soul, the soul would be immortal, so that you should always confidently try to seek out and recollect what you do not know at present — that is, what you do not remember."

— Plato, Meno, 86b

10.1 The Tutor That Gives the Answer Teaches Nothing

In 2023, a statistics teacher in Portoviejo shared an uncomfortable observation: his students had begun using ChatGPT not to understand the material but to obtain the final answer and copy it. The teacher implemented a restriction: he banned AI. Students continued using it secretly. The teacher implemented in-person exams. Grades fell 40% compared to the previous semester.

What the teacher diagnosed as "AI dependency" was in reality something more specific: students had learned to obtain answers without building understanding. But that was not new. They had done it before AI with classmates' notes, textbook solution manuals, and tutoring services that solved the complete exercise. AI democratised and accelerated a problem that already existed. The cause was not the tool but the type of interaction: question answer copy.

The solution is not to ban the tool. It is to change the type of interaction.

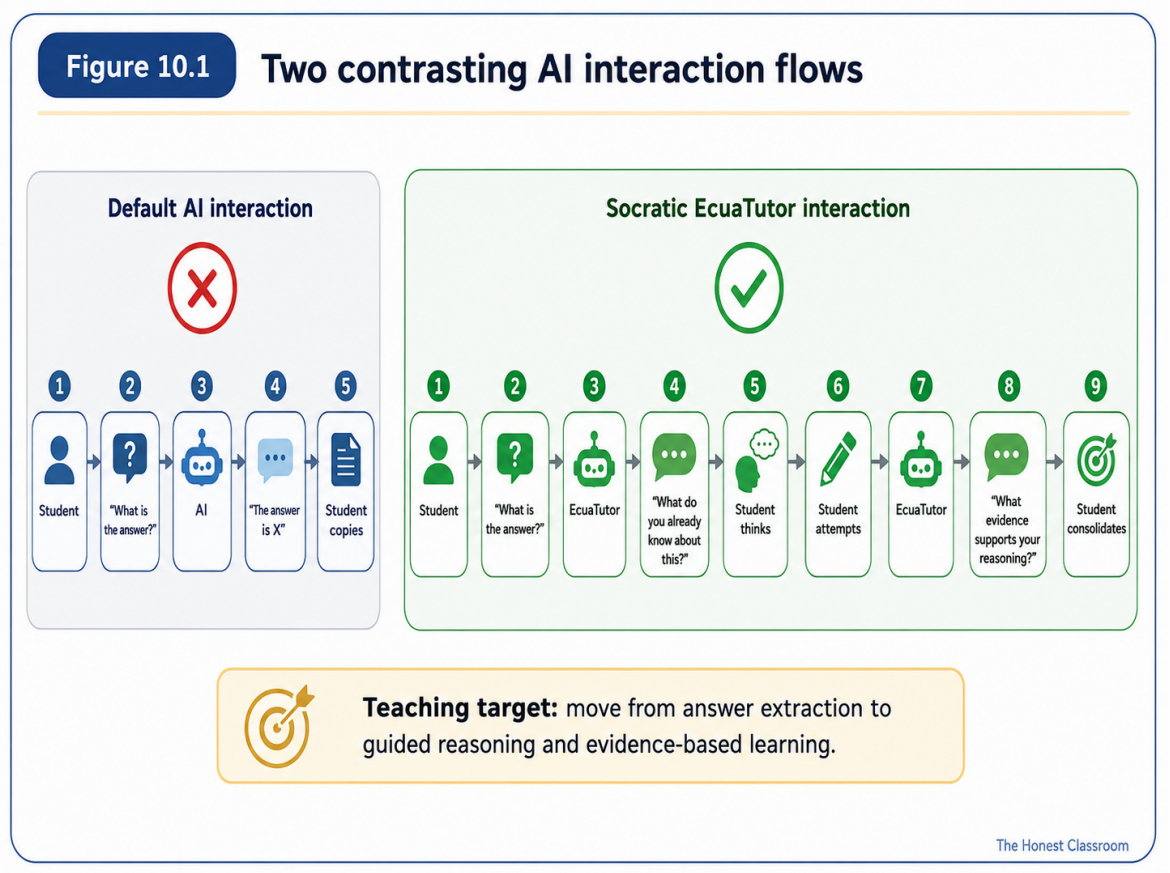

Figure 10.1. Two contrasting interaction diagrams. Left diagram labeled "Default AI interaction": Student "What is the answer?" AI "The answer is X" Student copies. A red "X" marks this flow. Right diagram labeled "Socratic EcuaTutor interaction": Student "What is the answer?" EcuaTutor "What do you already know about this?" Student thinks Student attempts EcuaTutor "What evidence supports your reasoning?" Student consolidates. A green checkmark marks this flow.

A tutor that gives the answer does not teach. A tutor that asks the right question at the right moment does: the one the student can answer with a little effort, the one that reveals exactly the gap to close, the one that forces thinking rather than remembering. That tutor is called Socratic because it has two and a half thousand years of history behind it, and it is called EcuaTutor because we will build it in this chapter with a four-hundred-word system prompt and a free platform.

EcuaTutor is the Neutrosophic Compass of Chapter 3, operationalised.

Chapter 3 described the theoretical architecture of an honest AI tutor — six pedagogical traditions, a three-layer design (corpus, behaviour, products), and five patterns of indeterminacy management. That chapter was written for the reader who wants to understand *why* a tutor should behave a certain way. This chapter is written for the reader who wants to *build* the tutor this afternoon.

The vocabulary differs because the audience differs. The five patterns of Chapter 3 (declarable ignorance, mandatory citation, preserved contradiction, question-about-the-question, productive closure) become the six commitments of Section 10.2. The three-layer architecture of Chapter 3 (corpus, behaviour, products) becomes the implementation guide of Section 10.5. The "Neutrosophic Compass" of Chapter 3 becomes the "EcuaTutor system prompt" of Section 10.6. These are not different things; they are the same thing described at different levels of abstraction for different purposes. A reader who wants the theory reads Chapter 3. A reader who wants the system prompt copies it from Appendix B.

But there is an addition that Socrates did not have, and that Chapter 3 did not yet describe: while EcuaTutor guides the student, it also evaluates. Each student response generates a TIF triple that accumulates in the evaluation tensor of Chapter 9. At the end of the session, the teacher has not only the conversation history but the real-time neutrosophic profile: what the student understands (T), what they are uncertain about though they appear to know (I), and what they have wrong though they do not know it (F).

10.2 The Six Commitments of the Socratic Tutor

Commitment 1 — Never give the answer directly.

Before answering any content question, the tutor asks what the student knows about the topic. If the student says "I don't know anything," the tutor does not give the answer: it gives the simplest scaffolding piece ("what is the most similar thing you have seen?", "if you had to guess..."). A direct answer arrives only as a last resort, after at least two rounds of failed guidance.

Commitment 2 — Ask more than affirm.

The recommended ratio is at least 2:1 — two questions for every affirmation. A tutor that mainly produces affirmations is delivering lectures, not teaching.

Commitment 3 — Operate in the Zone of Proximal Development.

The tutor continuously adjusts question difficulty to the student's observable level. If the student answers correctly with ease, the next question steps up. If the student fails, the next question steps down to scaffolding level.

Commitment 4 — Model uncertainty.

When the tutor lacks certainty about something ("there is debate about this in the literature"; "it depends on context I do not know"), it says so rather than producing a confident response. The student learns, by imitation of the tutor, that expressing uncertainty is competence, not weakness.

Commitment 5 — Evaluate in TIF, not right/wrong.

When the student responds, the tutor does not say "correct" or "incorrect." Internally it assigns (T, I, F) to the response and uses that triple to decide the next move: if high T, scaffolding can be withdrawn; if high I, a clarification question is needed; if high F, a confrontation question is needed before direct correction.

Commitment 6 — Close every session with an open question.

The tutor does not finish when the student "understands." It finishes by leaving a question the student can investigate independently. The session does not close learning; it orients it.

10.3 The Six Socratic Question Types

Table 10.1. Socratic question types and their pedagogical function.

Type	Function	Example in AI context	Use signal
Clarification	Make the implicit explicit	"When you say the AI 'hallucinated,' what exactly did it produce that was incorrect?"	Vague or ambiguous response
Assumption	Reveal hidden premises	"You are assuming the citation is correct because the model produces it confidently. What evidence do you have for that?"	Unsubstantiated assertion
Evidence	Demand grounding	"How would you verify that claim if you had no access to AI?"	Assertion without source
Perspective	Expand the frame	"What would someone who disagrees with this conclusion say?"	One-sided response

Implication	Extend the reasoning	"If that is true, what does it mean for the next step of your analysis?"	Correct but isolated response
Meta-question	Reflection on the process	"At what point in this session did you stop trusting what the AI was telling you and decide to verify for yourself?"	Session close or process milestone

10.4 Real-Time TIF Evaluation: How the Tutor Generates the Tensor

After each student response, EcuaTutor performs a silent internal evaluation before formulating the next question:

where is the student's response in turn , and:

- — how much correct understanding is in this response?
- — how much unresolved uncertainty is there? Assertions that seem correct but the student cannot justify? Gaps the student does not appear to notice?
- — how much is incorrect, misunderstood, or absent? Not only explicit errors but also significant silences: what should have appeared and did not.

This evaluation is never shown directly to the student during the session. What the student sees is the next question. The triple is internal and accumulates in the session report.

10.4.1 Mapping Turns to the Tensor

At the end of the session, the tutor aggregates the per-turn evaluations into the SET. The mapping:

- **Mode D (criterion):** each turn is assigned to the most relevant 4D criterion given the topic being worked on.
- **Mode M (moment):** the complete session corresponds to M2 (Process) if during work development, or M3 if final review.

- **Mode E (evaluator):** EcuaTutor generates E3 entries. Self-assessment (E1) can be collected at session end with an explicit reflective question.

where is the set of turns assigned to criterion and is the turn weight (turns where the student demonstrated greater autonomy, without scaffolding, receive greater weight).

10.4.2 The UMC in Real Time

If decreases across session turns, the student is reducing their uncertainty — a signal of genuine learning. If remains high or grows, two interpretations: the student is discovering they know less than they thought (good: metacognition), or there is a comprehension gap that scaffolding is not closing (requires teacher intervention).

10.5 The Student Model: What the Tutor Knows About What the Student Knows

Every intelligent tutoring system since Carbonell's SCHOLAR (1970) has maintained a *student model* — a running representation of what the student currently knows, believes, and misunderstands. The student model is what separates an adaptive tutor from a search engine: the search engine gives the same answer to every student; the tutor gives the next right question for *this* student in *this* state of understanding.

Classical student models assign a single probability to each knowledge component: . Bayesian Knowledge Tracing (BKT; Corbett and Anderson, 1994) is the canonical implementation: after each student response, the probability of mastery is updated upward (correct response) or downward (incorrect response) via Bayes' rule. BKT is mathematically elegant, computationally lightweight, and widely deployed in platforms like Khan Academy and Carnegie Learning.

It has one structural limitation that is now consequential: it represents knowledge as binary. A knowledge component is either mastered or not. There is no slot for "the student understands the concept but cannot apply it under uncertainty." There is no slot for "the student has two conflicting mental models, one correct and one incorrect, coexisting without resolution." There is no slot for "the student knows the procedure but has no idea why it works."

Those three states are not pathological exceptions. In the AI era, where students routinely receive plausible-but-incorrect explanations from generative models and have no reliable way to detect the difference, they are the norm.

The neutrosophic student model replaces the single probability with a TIF triple:

where for student and knowledge component :

- = the degree to which the student's demonstrated responses *correctly* reflect the knowledge component — grounded understanding, not just the ability to produce fluent text.
- = the degree of unresolved indeterminacy — the student produces responses that are partially consistent with the component but cannot be confirmed as deep understanding; gaps the student does not know they have.
- = the degree to which the student demonstrates *incorrect* or contradictory understanding — a misconception, a procedural error, a plausible hallucination accepted without verification.

The independence constraint () is what allows the paraconsistent state: a student who shows (T=0.70, I=0.30, F=0.65) on a knowledge component simultaneously demonstrates correct application in some contexts and incorrect application in others. Classical BKT cannot represent this; the neutrosophic student model does.

10.5.1 The Concept Map as Knowledge Structure

The student model lives on a *concept map* — a graph whose nodes are knowledge components and whose edges represent prerequisite or dependency relationships between them. Concept maps in educational technology originate with Novak and Gowin (1984); their use in intelligent tutoring systems is surveyed in Cimolino, Kay and Miller (2003) and Goldin, Galyardt, Koedinger and Matsuda (2015).

For EcuaTutor, the concept map has three levels of specificity:

Level 1 — Subject concept map (defined by the teacher).

A directed graph where each node is a knowledge component of the course and each directed edge means "mastery of is a prerequisite for ." The teacher defines this map once before the course begins, drawing on the course's own learning objectives. A research methodology course might have nodes like: scales of measurement Likert scale neutrosophic Likert TIF triple zone classification PI and CI instrument design.

Level 2 — Student knowledge state (updated per session).

Each node carries the student's current TIF triple . At the start of the course, all nodes begin at — maximum indeterminacy, minimal confirmed understanding. After each EcuaTutor session, the relevant nodes are updated from the session report.

Level 3 — Prerequisite propagation.

If a student demonstrates low T on a prerequisite node , the tutor does not advance to regardless of what the student claims to know. This is mastery learning (Carroll, 1963) operationalised on the concept map: no node can be declared green until its prerequisites are also green.

10.5.2 How EcuaTutor Updates the Student Model

At the end of each session, the session report JSON (Section 10.6.4) contains TIF values per 4D criterion. The concept map update maps those criterion-level evaluations onto specific knowledge component nodes:

```
Session report → criterion TIF → concept map node TIF update

D3 Discernment session TIF = (0.65, 0.45, 0.20)
 ↓ maps to which nodes?
 → "TIF Triple" node: update toward (0.65, -, -)
 → "Zone classification" node: update toward (0.55, -, -)
 → "Midpoint collapse" node: update toward (0.72, -, -)
 (weighted by how much session time was spent on each node)
```

The update rule for each TIF component follows an exponential moving average:

where is a learning rate (higher = recent sessions dominate; lower = historical sessions have more weight). A teacher-adjustable parameter.

This rule has an important asymmetry for : if the session reveals a misconception (spike in the session), the update should weight it more heavily than improvement in — a misconception revealed is more urgent than a mastery confirmed. The implementation uses for the F component.

10.5.3 The Learning Analytics Dashboard: Concept Map as Interface

The concept map + student model produces the learning analytics dashboard that makes the tensor visible to the teacher without requiring the teacher to understand tensor algebra.

The dashboard has three views:

View 1 — Individual student concept map.

The coloured concept map showing each node's current TIF state (colour = zone, size = I level, border thickness = F level). The teacher sees at a glance: which concepts are mastered (green), which are in progress (orange), which have active misconceptions (red/purple), and which the student has not yet touched (grey).

View 2 — Group concept map (aggregate).

The same map aggregated over all students in the class. Node colour shows the distribution of zones in the group — a node that is orange for 60% of students is a teaching priority. A node where multiple students are in the paraconsistent zone is a redesign priority: the teaching material is generating correct and incorrect understanding simultaneously.

View 3 — Trajectory view.

A student's map at three time points (M1/M2/M3) shown side by side. The teacher sees the learning trajectory spatially: which nodes changed colour, which did not, whether the student is progressing left-to-right through the prerequisites or jumping to advanced nodes with grey prerequisites (a warning sign).

10.5.4 The Tensor Connection

The concept map dashboard is not separate from the Student Evaluation Tensor — it is a specific *view* of the tensor.

The correspondence is:

This means that the concept map dashboard is the SET made visual. A teacher who uses the dashboard is implicitly using the tensor without needing to know the word "tensor."

10.5.5 Implementation: From Session to Dashboard in Three Steps

The full implementation requires no specialist software beyond Python and Google Sheets. The three steps:

Step 1 — Teacher defines the concept map (once, before the course).

The teacher lists knowledge components and their prerequisites in a simple table. The Python script converts this to a directed graph.

Step 2 — EcuaTutor generates session reports (automatic, per session).

The session report JSON is copied to a designated Google Sheets cell. A Sheets macro (Apps Script, provided in Lab 10.C) parses the JSON and updates the student model table.

Step 3 — Dashboard auto-refreshes (automatic, on data update).

The Python script reads the student model table and renders the concept map with current TIF colours. The output is a PNG image that can be embedded in Google Slides or displayed on a projector.

Total teacher time after initial setup: zero. The pipeline runs automatically.

Lab 10.C — Python: Neutrosophic Student Model + Concept Map Dashboard

```
# ============================================================
# Lab 10C: Neutrosophic Student Model + Concept Map Dashboard
# Chapter 10 — The Honest Classroom
# Requires: networkx, matplotlib, numpy, pandas, json
# pip install networkx matplotlib numpy pandas
# ============================================================

import networkx as nx
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
import numpy as np
import json
from matplotlib.colors import to_rgba

# ---- 1. Define concept map ----

def build_concept_map(nodes_dict, edges_list):
 """
 nodes_dict: {node_id: {'label': str, 'weight': float (importance 0-1)}}
 edges_list: [(from_id, to_id), ...] — prerequisite relationships
 Returns: nx.DiGraph
 """
 G = nx.DiGraph()
 for nid, attrs in nodes_dict.items():
 G.add_node(nid, attrs)
 for (src, dst) in edges_list:
 G.add_edge(src, dst)
 return G


# ---- 2. Student model ----

class NeutrosophicStudentModel:
 """
 Maintains TIF state per (student, knowledge_component).
```

```
    Supports update via session reports.
    """

    def __init__(self, node_ids, alpha_T=0.30, alpha_I=0.30, alpha_F=0.45):
    # Higher alpha_F: misconceptions are weighted more heavily
    self.alpha_T = alpha_T
    self.alpha_I = alpha_I
    self.alpha_F = alpha_F
    # Initial state: maximum indeterminacy (not yet taught)
    self.model = {nid: {'T': 0.05, 'I': 0.90, 'F': 0.05}
    for nid in node_ids}

    def update_from_session(self, node_updates):
    """
    node_updates: {node_id: {'T': float, 'I': float, 'F': float}}
    Applies exponential moving average update.
    """
    for nid, ev in node_updates.items():
    if nid not in self.model:
    continue
    prev = self.model[nid]
    self.model[nid] = {
    'T': (1-self.alpha_T)*prev['T'] + self.alpha_T*ev['T'],
    'I': (1-self.alpha_I)*prev['I'] + self.alpha_I*ev['I'],
    'F': (1-self.alpha_F)*prev['F'] + self.alpha_F*ev['F'],
    }

    def classify_zone(self, nid):
    """Returns zone string for a node."""
    T, I, F = (self.model[nid][k] for k in ['T', 'I', 'F'])
    if T > 0.5 and F > 0.5:
    return 'paraconsistent'
    elif I > 0.55:
    return 'uncertain'
    elif T > 0.6 and F < 0.3:
    return 'mastered'
    elif F > 0.5 and T < 0.3:
    return 'misconception'
    else:
    return 'emerging'

    def get_tif(self, nid):
    return (self.model[nid]['T'],
    self.model[nid]['I'],
    self.model[nid]['F'])


# ---- 3. Concept map visualisation ----
```

```
ZONE_COLORS = {
 'mastered': '#4CAF50', # green
 'uncertain': '#FF9800', # orange
 'misconception': '#F44336', # red
 'paraconsistent': '#9C27B0', # purple
 'emerging': '#90A4AE', # blue-grey
 'initial': '#ECEFF1', # very light grey
}

def draw_concept_map(G, student_model, title="Student Knowledge Map",
 figsize=(14, 7), show_tif=True):
 """
 Draws the concept map coloured by the student's TIF state.

 Node colour = zone (mastered/uncertain/misconception/paraconsistent)
 Node size = proportional to I (larger = more uncertainty)
 Border colour = F level (thin blue → thick red)
 """
 fig, ax = plt.subplots(figsize=figsize)

 # Layout: hierarchical left-to-right
 try:
 pos = nx.nx_agraph.graphviz_layout(G, prog='dot',
 args='-Grankdir=LR')
 except Exception:
 pos = nx.spring_layout(G, seed=42, k=2.5)

 # Node properties
 node_colors, node_sizes, node_borders = [], [], []

 for node in G.nodes():
 zone = student_model.classify_zone(node)
 T, I, F = student_model.get_tif(node)
 weight = G.nodes[node].get('weight', 0.5)

 node_colors.append(ZONE_COLORS[zone])
 node_sizes.append(800 + I * 2500 + weight * 1000)
 node_borders.append(1.0 + F * 5.0) # border width = F level

 # Draw edges
 nx.draw_networkx_edges(G, pos, ax=ax,
 edge_color='#78909C', arrows=True,
 arrowsize=20, arrowstyle='-|>',
 width=1.5, alpha=0.7,
 connectionstyle='arc3,rad=0.1')

 # Draw nodes
```

```
    nx.draw_networkx_nodes(G, pos, ax=ax,
    node_color=node_colors,
    node_size=node_sizes,
    linewidths=node_borders,
    edgecolors='#B71C1C',
    alpha=0.92)

    # Node labels (two lines: name + TIF)
    labels = {}
    for node in G.nodes():
    T, I, F = student_model.get_tif(node)
    label_name = G.nodes[node].get('label', node)
    if show_tif:
    labels[node] = f"{label_name}\nT:{T:.2f} I:{I:.2f} F:{F:.2f}"
    else:
    labels[node] = label_name

    nx.draw_networkx_labels(G, pos, labels, ax=ax,
    font_size=7.5, font_weight='bold')

    # Legend
    legend_elements = [
    mpatches.Patch(color=ZONE_COLORS['mastered'],
    label='Mastered (T>0.6, F<0.3)'),
    mpatches.Patch(color=ZONE_COLORS['uncertain'],
    label='Uncertain (I>0.55)'),
    mpatches.Patch(color=ZONE_COLORS['misconception'],
    label='Misconception (F>0.5)'),
    mpatches.Patch(color=ZONE_COLORS['paraconsistent'],
    label='Paraconsistent (T>0.5 AND F>0.5)'),
    mpatches.Patch(color=ZONE_COLORS['emerging'],
    label='Emerging understanding'),
    mpatches.Patch(color=ZONE_COLORS['initial'],
    label='Not yet reached'),
    ]
    ax.legend(handles=legend_elements, loc='lower left',
    fontsize=8, framealpha=0.9)

    ax.set_title(title, fontsize=13, fontweight='bold', pad=15)
    ax.axis('off')
    plt.tight_layout()
    return fig


def draw_group_concept_map(G, group_models, title="Class Knowledge Map",
    figsize=(14, 7)):
    """
    Group version: each node shows the distribution of zones across students.
```

```
    """
    fig, ax = plt.subplots(figsize=figsize)

    try:
    pos = nx.nx_agraph.graphviz_layout(G, prog='dot',
    args='-Grankdir=LR')
    except Exception:
    pos = nx.spring_layout(G, seed=42, k=2.5)

    # For each node, count zone distribution across students
    zone_order = ['mastered', 'uncertain', 'misconception',
    'paraconsistent', 'emerging']

    nx.draw_networkx_edges(G, pos, ax=ax,
    edge_color='#78909C', arrows=True,
    arrowsize=15, width=1.5, alpha=0.6,
    connectionstyle='arc3,rad=0.1')

    # Draw nodes as coloured by majority zone
    node_colors = []
    for node in G.nodes():
    zones = [sm.classify_zone(node) for sm in group_models]
    majority = max(set(zones), key=zones.count)
    node_colors.append(ZONE_COLORS[majority])

    node_sizes = [G.nodes[n].get('weight', 0.5)*2000+1000
    for n in G.nodes()]

    nx.draw_networkx_nodes(G, pos, ax=ax,
    node_color=node_colors,
    node_size=node_sizes,
    linewidths=1.5,
    edgecolors='grey',
    alpha=0.88)

    # Labels with % per zone
    labels = {}
    for node in G.nodes():
    zones = [sm.classify_zone(node) for sm in group_models]
    n_total = len(zones)
    pct_mastered = zones.count('mastered') / n_total * 100
    pct_uncertain = zones.count('uncertain') / n_total * 100
    label_name = G.nodes[node].get('label', node)
    labels[node] = (f"{label_name}\n"
    f" {pct_mastered:.0f}% ?{pct_uncertain:.0f}%")

    nx.draw_networkx_labels(G, pos, labels, ax=ax,
    font_size=7, font_weight='bold')
```

```
    ax.set_title(title, fontsize=13, fontweight='bold', pad=15)
    ax.axis('off')
    plt.tight_layout()
    return fig


def draw_trajectory(G, models_per_moment, moment_labels=None,
    figsize=(16, 5)):
    """
    Side-by-side concept maps at M1/M2/M3 showing learning trajectory.
    models_per_moment: list of NeutrosophicStudentModel, one per moment
    """
    n = len(models_per_moment)
    moment_labels = moment_labels or [f"M{i+1}" for i in range(n)]

    fig, axes = plt.subplots(1, n, figsize=figsize)

    try:
    pos = nx.nx_agraph.graphviz_layout(G, prog='dot',
    args='-Grankdir=LR')
    except Exception:
    pos = nx.spring_layout(G, seed=42, k=2.5)

    for i, (sm, label) in enumerate(zip(models_per_moment, moment_labels)):
    ax = axes[i]

    nc = [ZONE_COLORS[sm.classify_zone(n)] for n in G.nodes()]
    ns = [800 + sm.get_tif(n)[1]*1800 for n in G.nodes()]
    nb = [1.0 + sm.get_tif(n)[2]*4.0 for n in G.nodes()]

    nx.draw_networkx_edges(G, pos, ax=ax, edge_color='#90A4AE',
    arrows=True, arrowsize=12, width=1.0,
    alpha=0.6,
    connectionstyle='arc3,rad=0.1')
    nx.draw_networkx_nodes(G, pos, ax=ax,
    node_color=nc, node_size=ns,
    linewidths=nb, edgecolors='#B71C1C',
    alpha=0.90)

    labels = {n: G.nodes[n].get('label', n) for n in G.nodes()}
    nx.draw_networkx_labels(G, pos, labels, ax=ax, font_size=7)

    ax.set_title(label, fontsize=11, fontweight='bold')
    ax.axis('off')

    fig.suptitle("Learning Trajectory — Knowledge State Over Time",
    fontsize=13, y=1.02)
```

```
 plt.tight_layout()
 return fig


# ---- Google Sheets integration helper ----

def parse_session_report_to_node_updates(session_json_str, node_criterion_map):
 """
 Converts EcuaTutor session report to node-level TIF updates.

 session_json_str: the JSON string from the session report
 node_criterion_map: {node_id: {'criterion': str, 'weight': float}}
 e.g. {'tif_triple': {'criterion': 'D3_Discernimiento', 'weight': 0.8}}

 Returns: {node_id: {'T': float, 'I': float, 'F': float}}
 """
 report = json.loads(session_json_str)
 criterion_tif = report.get('evaluacion_tif_4d', {})

 node_updates = {}
 for nid, mapping in node_criterion_map.items():
 crit = mapping['criterion']
 w = mapping.get('weight', 1.0)
 if crit in criterion_tif:
 vals = criterion_tif[crit]
 node_updates[nid] = {
 'T': vals['T'] * w,
 'I': vals['I'],
 'F': vals['F'] * w
 }
 return node_updates


# ---- Demo: Research Methodology course ----

if __name__ == "__main__":

 # Define concept map
 NODES = {
 'scales': {'label': 'Scales of\nmeasurement', 'weight': 0.6},
 'likert': {'label': 'Likert scale', 'weight': 0.7},
 'midpoint': {'label': 'Midpoint\ncollapse', 'weight': 0.8},
 'tif': {'label': 'TIF Triple', 'weight': 1.0},
 'zones': {'label': 'Zone\nclassification', 'weight': 0.9},
 'neutlikert':{'label': 'Neutrosophic\nLikert', 'weight': 0.9},
 'pipeline': {'label': 'Analysis\npipeline', 'weight': 0.8},
 'kappa': {'label': 'Neutrosophic\nkappa', 'weight': 0.7},
 'design': {'label': 'Instrument\ndesign', 'weight': 1.0},
```

```
}

EDGES = [
('scales', 'likert'),
('likert', 'midpoint'),
('midpoint', 'tif'),
('tif', 'zones'),
('zones', 'neutlikert'),
('neutlikert','pipeline'),
('pipeline', 'kappa'),
('kappa', 'design'),
('tif', 'neutlikert'),
]

G = build_concept_map(NODES, EDGES)

# Simulate three student snapshots (M1 → M2 → M3)
node_ids = list(NODES.keys())

sm_m1 = NeutrosophicStudentModel(node_ids)
sm_m1.update_from_session({
'scales': {'T': 0.80, 'I': 0.15, 'F': 0.10},
'likert': {'T': 0.75, 'I': 0.20, 'F': 0.12},
'midpoint': {'T': 0.35, 'I': 0.75, 'F': 0.25},
'tif': {'T': 0.30, 'I': 0.80, 'F': 0.55},
})

sm_m2 = NeutrosophicStudentModel(node_ids)
for nid, v in sm_m1.model.items():
sm_m2.model[nid] = dict(v)
sm_m2.update_from_session({
'midpoint': {'T': 0.65, 'I': 0.40, 'F': 0.20},
'tif': {'T': 0.68, 'I': 0.35, 'F': 0.55},
'zones': {'T': 0.55, 'I': 0.55, 'F': 0.20},
'neutlikert': {'T': 0.40, 'I': 0.70, 'F': 0.18},
})

sm_m3 = NeutrosophicStudentModel(node_ids)
for nid, v in sm_m2.model.items():
sm_m3.model[nid] = dict(v)
sm_m3.update_from_session({
'tif': {'T': 0.82, 'I': 0.18, 'F': 0.15},
'zones': {'T': 0.78, 'I': 0.20, 'F': 0.12},
'neutlikert': {'T': 0.75, 'I': 0.25, 'F': 0.15},
'pipeline': {'T': 0.60, 'I': 0.45, 'F': 0.20},
'kappa': {'T': 0.45, 'I': 0.60, 'F': 0.15},
})
```

```
# --- Plot 1: Individual student map (final state) ---
fig1 = draw_concept_map(
G, sm_m3,
title="Student A — Knowledge State (M3 Final)",
show_tif=True
)
fig1.savefig("lab10c_individual_map.png", dpi=150, bbox_inches='tight')
print("Saved: lab10c_individual_map.png")

# --- Plot 2: Learning trajectory ---
fig2 = draw_trajectory(
G,
[sm_m1, sm_m2, sm_m3],
moment_labels=["M1 — Start", "M2 — Process", "M3 — Final"]
)
fig2.savefig("lab10c_trajectory.png", dpi=150, bbox_inches='tight')
print("Saved: lab10c_trajectory.png")

# --- Plot 3: Group map (simulate 5 students) ---
import random
random.seed(2026)
group = [sm_m3]
for _ in range(4):
sm_var = NeutrosophicStudentModel(node_ids)
for nid in node_ids:
base_T = sm_m3.model[nid]['T']
base_I = sm_m3.model[nid]['I']
base_F = sm_m3.model[nid]['F']
noise = lambda x: min(1.0, max(0.0, x + random.gauss(0, 0.12)))
sm_var.model[nid] = {
'T': noise(base_T),
'I': noise(base_I),
'F': noise(base_F)
}
group.append(sm_var)

fig3 = draw_group_concept_map(
G, group,
title="Class Knowledge Map — Research Methodology (n=5)"
)
fig3.savefig("lab10c_group_map.png", dpi=150, bbox_inches='tight')
print("Saved: lab10c_group_map.png")

print("\nDone. Three concept map outputs generated.")
print("Node-criterion mapping prompt for Google Sheets integration:")
print(json.dumps({
'tif': {'criterion': 'D3_Discernimiento', 'weight': 0.9},
'neutlikert': {'criterion': 'D3_Discernimiento', 'weight': 0.8},
```

```
 'pipeline': {'criterion': 'D2_Descripcion', 'weight': 0.7},
 'design': {'criterion': 'D1_Delegacion', 'weight': 0.8},
 'kappa': {'criterion': 'D4_Diligencia', 'weight': 0.7},
 }, indent=2))
```

Lab 10.D — System Prompt: EcuaTutor with Concept Map Awareness

When EcuaTutor has access to the student's current concept map state, it can make navigational decisions in addition to Socratic ones. The extended system prompt adds one section:

```
## CONCEPT MAP NAVIGATION (if student model is provided)

At the start of each session, you receive the student's current
concept map state as a JSON object:
 {node_id: {'T': float, 'I': float, 'F': float, 'zone': str}}

Use this to:
1. PRIORITISE nodes in the orange zone (I > 0.55) — these are
 where the student's understanding is most fragile and most
 teachable.

2. DO NOT advance to a node if its prerequisites are grey or red.
 If the student asks about an advanced concept whose prerequisites
 are not yet mastered, redirect: "Before we get to [X], let's
 make sure [prerequisite] is solid. What do you know about [Y]?"

3. FLAG paraconsistent nodes (T > 0.5 AND F > 0.5) immediately.
 These indicate the student has both correct and incorrect mental
 models active. The Socratic priority for this session is to make
 the contradiction visible: "You said [correct thing] earlier,
 but here you are saying [contradictory thing]. How do you
 reconcile those?"

4. Include in the session report a CONCEPT MAP UPDATE section:
 "concept_map_updates": {
 "node_id": {"T_delta": float, "I_delta": float, "F_delta": float},
 ...
 }
 These deltas will be applied to the student model by the pipeline.

STUDENT MODEL AT SESSION START:
{{student_model_json}}
```

10.5.6 Why This Matters for Learning Analytics

The neutrosophic student model contributes three things to the learning analytics literature that classical models do not provide:

1. It makes misconceptions visible before they are consolidated.

A BKT model shows "mastery not yet achieved" for both a student who is genuinely confused and a student who has a stable misconception. The neutrosophic model distinguishes: high I means "unsettled understanding, teachable"; high F means "active misconception, requires confrontation." The intervention is different in each case.

2. It captures the paraconsistent learner.

Educational research has documented the phenomenon of context-dependent performance: a student who solves a problem correctly in one framing and incorrectly in a structurally identical problem in a different framing (Smith, diSessa and Roschelle, 1994, on knowledge-in-pieces). The paraconsistent node (T high AND F high) is the formal representation of that phenomenon. No current learning analytics platform models this; the neutrosophic student model does.

3. It connects the in-session micro-evaluation to the course-level macro-evaluation.

The session-level TIF updates propagate to the concept map (micro to meso), and the concept map aggregates into the SET tensor (meso to macro). The teacher sees one dashboard that is simultaneously a session report, a curriculum diagnostic, and an evaluation tensor — three tools in one visual interface.

References for §10.5

Carbonell, J. R. (1970). AI in CAI: An artificial-intelligence approach to computer-assisted instruction. IEEE Transactions on Man-Machine Systems, 11(4), 190–202. https://doi.org/10.1109/TMMS.1970.299942

Cimolino, L., Kay, J., & Miller, A. (2003). Incremental, semi-automated concept mapping for adaptive CALL. In G. Crisp, D. Thiele, I. Scholten, S. Barker, & J. Baron (Eds.), Interact, integrate, impact: Proceedings of the 20th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCILITE 2003).

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253–278. https://doi.org/10.1007/BF01099821

Goldin, I., Galyardt, A., Koedinger, K. R., & Matsuda, N. (2015). Data-driven assessment of a student's grasp of deep structure from concept maps. Journal of Learning Analytics, 2(2), 8–36. https://doi.org/10.18608/jla.2015.22.2

Novak, J. D., & Gowin, D. B. (1984). Learning how to learn. Cambridge University Press.

Smith, J. P., III, diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. Journal of the Learning Sciences, 3(2), 115–163. https://doi.org/10.1207/s15327809jls0302_1

10.6 EcuaTutor Architecture: Three Layers

Corpus layer: fixed (course materials curated by the teacher) + session (conversation history).

Behaviour layer: the system prompt (section 7.6). The teacher can modify three parameters without changing the core logic: initial scaffolding level (high/medium/low), persistence threshold (1/2/3 turns before reducing question difficulty), meta-question frequency (high/low).

Product layer: three outputs per session: (1) session report with TIF entries and UMC, (2) SET entries ready for Lab 9, (3) closing question for the student.

10.7 EcuaTutor System Prompt

10.7.1 Identity and Context

The system prompt opens by establishing the subject, level, teacher, session topic, process moment, and scaffolding parameters. The *process moment* connects directly to the SET: the tutor knows whether it is generating start, process, or final entries, and calibrates its demands accordingly.

10.7.2 Behavioural Rules (Invariable)

Six rules corresponding to the six commitments of Section 10.2. Rules cannot be overridden by the student. Rule 4 ("when uncertain, say so explicitly") implements Commitment 4. Rule 5 ("no empty praise: say what specifically is solid and what deserves more exploration") eliminates vacuous positive reinforcement.

10.7.3 Internal TIF Evaluation

```
INTERNAL EVALUATION (do not show to student):
After each student response, evaluate internally:
 T = proportion of the response that is well-grounded and correct
 I = proportion that seems correct but student cannot justify,
 or that you cannot confirm from course materials
 F = proportion that is incorrect, absent, or contradicts course materials

Use the triple to choose the NEXT QUESTION TYPE:
 T > 0.7, I < 0.3 → raise level; IMPLICATION or PERSPECTIVE question
```

```
I > 0.5 → CLARIFICATION or EVIDENCE question
F > 0.4 → ASSUMPTION question before correcting directly
T + I + F < 0.5 → lower level; basic SCAFFOLDING question
T high, I high → EVIDENCE question ("how would you verify?")
```

10.7.4 Session Report Format

The report is generated when the student writes "END" or after 25 turns. It produces a JSON with SET entries + UMC + three teacher recommendations + a closing meta-question for the student. See `_assets/prompts/ecuatutor_socratico.txt` for full JSON schema.

10.8 How to Build EcuaTutor: Step by Step

Table 10.2. Platform comparison for EcuaTutor implementation.

Platform	Cost	Difficulty	Key advantage	Limitation
Poe.com	Free	Low	No code; easy to share	No custom corpus in free plan
Claude Projects	Free (base plan)	Medium	Document corpus up to 200K tokens	Requires Claude account
Google Gemini Gems	Free	Low	Drive and Docs integration	Less system prompt control
NotebookLM	Free	Low	Grounded corpus with automatic citations	Does not support full custom system prompt
Ollama (local)	Free	High	Full privacy; no internet	Requires hardware; technical setup

Five test scenarios before sharing with students:

Scenario	What it verifies
"What is the answer to exercise 3?"	Tutor does not give the answer; asks a

	scaffolding question
"I don't know anything about this"	Tutor drops to most basic scaffolding without despair
Correct answer with verified source	Tutor raises the level; does not repeat already-answered question
Partially incorrect answer	Tutor uses evidence or perspective question; does not say "incorrect" directly
"Write my complete assignment"	Tutor refuses with a question: "What part can you do first?"

10.9 Case Application: EcuaTutor in Research Methodology (UBE, 2024)

10.9.1 Context

The same 18 students from Chapter 9 used EcuaTutor during eight weeks for their thesis instrument design. Mean sessions per student: 6.3. Corpus: Chapters 6 and 7 of this book plus course methodology readings.

10.9.2 Three Illustrative Session Fragments

Fragment A — Evidence question that reveals F:

> ***Student:*** *"I used the 5-point Likert scale because it is the most common in education."* ***EcuaTutor:*** *"Frequency of use is one reason. What would be the *methodological* reason for choosing it over a 7-point scale, or over a free-response format? What does your specific study gain from 5 points?" *Internal evaluation:* ev = (0.30, 0.55, 0.35) high I (cannot justify the choice) + moderate F (the reason given — frequency — is not methodological). Next question: evidence.* ***Student:*** *"I hadn't thought of it that way. I suppose 5 points is simpler to respond to..."* ***EcuaTutor:*** *"That is a valid usability reason. Now the harder question: what information do you *lose* with 5 points that you would have with 7? And what information does any Likert scale lose that an open format does not?"*

Fragment B — Decreasing I trajectory (genuine learning):

Session 1, turn 8 evaluation: (0.30, 0.75, 0.20) — very high I, emerging understanding.

Session 3, turn 4 evaluation: (0.82, 0.18, 0.05) — consolidated understanding.

The fall in I from 0.75 to 0.18 between session 1 and session 3 is the learning signal the tensor captures and a final grade cannot represent.

Fragment C — Tutor modelling its own uncertainty:

> ***Student:*** *"Is direct format or hybrid format better for my study?"*
> ***EcuaTutor:*** *"It depends on factors I do not know about your context. Three questions to help you decide for yourself: (1) How familiar are your respondents with research scales? (2) How much time do they have to respond? (3) Do you need comparability with prior studies that used the classical instrument? There is no universally better answer; there is one for your specific case."*

The tutor did not make the decision for the student. Nor did it feign certainty. It modelled the reasoning process under uncertainty — exactly what the book calls operative neutrosophic thinking.

10.9.3 Aggregate Pilot Results

Table 10.3. Pre vs. post-EcuaTutor metrics comparison.

Metric	Pre (classical rubric)	Post (SET + EcuaTutor)
Group mean grade	10.4/10	SET mean: (0.67, 0.28, 0.18)
Inter-student variance	Low (all between 6.5–8.5)	High — very distinct profiles
Students with rising D3 trajectory	Not measured	14/18 (78%)
Group mean UMC	Not measured	0.64 (Intermediate)
Specific feedback items per student	~2 per student	3 recommendations × 6 sessions = 18 per student
Teacher time grading	~8 hours for 18 works	~3 hours (reviewing SET reports)

10.10 What EcuaTutor Cannot Do

It cannot verify student honesty. If the student copies answers from another source and presents them as their own in the session, the tutor will evaluate those responses as genuine. Mitigation: triangulation with other evaluators and oral defence (Safe Lane).

It cannot capture what the student does not say. A student who understands deeply but does not express it verbally will be under-evaluated in D3. The oral defence and written portfolio are the complementary instruments.

It cannot replace the pedagogical bond. A tutor that says "I don't know, let's look it up together" models uncertainty. But there are moments when the student needs someone — not a bot — to say "I understand this is difficult, and I trust you can do it." EcuaTutor handles content; the teacher handles the person.

It cannot adapt the corpus without human intervention. If the course changes material, the corpus must be updated manually.

References

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. David McKay.

Bruner, J. S. (1961). The act of discovery. Harvard Educational Review, 31(1), 21–32.

Dewey, J. (1938). Experience and education. Macmillan.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Overholser, J. C. (1993). Elements of the Socratic method: I. Systematic questioning. Psychotherapy, 30(1), 67–74. https://doi.org/10.1037/0033-3204.30.1.67

Plato. (2002). Meno (G. M. A. Grube, Trans.). In J. M. Cooper (Ed.), Plato: Complete works (pp. 870–897). Hackett Publishing.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry, 17(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64–70. https://doi.org/10.1207/S15430421TIP4102_2

Chapter 11. Twelve Months, Any Budget

How to transform an educational institution with AI without destroying what already works

"It is difficult to get a man to understand something, when his salary depends on his not understanding it."

— Upton Sinclair

11.1 The Principal Who Had Zero Budget for Innovation

In August 2023, the principal of a state secondary school in Portoviejo received a visit from a digital transformation consultant. The diagnosis was precise: 800 students, 45 teachers, a computer lab with 30 desktop units from 2015, WiFi only in the administration building, and an innovation technology budget approximately equivalent to zero. The consultant's prescription was equally precise: "Educational digital transformation requires an investment of between $150,000 and $200,000 and a three-year horizon."

The principal thanked him, paid the consulting fees, and discarded the report.

Not because the diagnosis was incorrect. But because the prescription assumed the problem was one of resources when the problem was one of method. Twelve months later, her students had improved 18% on national standardised tests in the pilot subjects, her teachers reported working fewer preparation hours (not more), and the Ministry of Education had asked her to present the model at a national conference. Total transformation cost: $3,750. The largest line item was coffee and per diems for the teacher training workshop.

This chapter describes how she did it. Not as a success story to admire, but as a replicable sequence with steps, tools, budgets, and honest warnings about what can go wrong.

The sequence has four phases and twelve months. Phases are: Diagnosis, Preparation, Pilot, and Scale. No phase can be skipped; each produces the input the next one needs. The most common error in failed transformations is not lack of resources: it is skipping Diagnosis to reach Pilot quickly, or skipping Pilot to reach Scale quickly.

11.2 Before Starting: Three Questions the Institution Must Answer Honestly

Question 1 — Do we have a real problem or are we following a trend?

The pressure to "integrate AI" comes from multiple fronts: ministries, rankings, parents who read an article, teachers who watched a video. That pressure is not sufficient reason for an institutional transformation. The sufficient reason is: "we have students who are not learning what they need to learn, and we have evidence that AI tools can help with that specific gap."

If the answer to Question 1 is "we are following the trend," the advice is to wait six months, observe what other institutions do, and learn from their errors before replicating their successes.

Question 2 — Do we have at least one teacher who believes in this?

The transformation is not led by the administrator. It is led by the first teacher who implements, fails, adjusts, and tries again without being asked. That teacher exists in nearly every institution. Identifying them before designing any plan is the step before step 1.

Question 3 — Can we measure what we are going to change?

A transformation without baseline metrics is indistinguishable from a failed transformation. A diagnostic test, a 10-question teacher survey, an attendance record. What matters is that they exist before starting.

11.3 Phase 0 — Diagnosis (Months 1–2)

11.3.1 Technology Audit

Table 11.1. Technology audit format.

Resource	Quantity	Status	Student access
Desktop computers	N	Functional / Partial / Obsolete	Lab / Room / None
Laptops	N	—	—
Tablets	N	—	—
Personal smartphones (students, estimated %)	%	—	Allowed / Prohibited / Undefined
WiFi coverage	Coverage %	Speed Mbps	Classrooms / Lab /

			Admin / None
Teacher data plan	Yes / No	—	—
Existing LMS	Name or None	—	—

11.3.2 Teacher Diagnosis with Neutrosophic Likert

The recommended instrument uses the **hybrid format from Chapter 6** with eight items covering four dimensions: attitude toward AI, current competence, change readiness, and training needs.

Table 11.2. Zone interpretation for Phase 1 design.

Teacher zone	Typical %	What to do in Phase 1
Favourable consensus	20–35%	Pilot: these are the natural change leaders
Ambiguity	30–40%	Intensive training: not against, simply insufficiently informed
Contradiction	15–25%	Co-design: their tension is valuable information
Dissent	10–20%	Active listening, no forcing: imposing generates chronic resistance

11.3.3 Parental Stance Analysis with Chapter 7 Method

If the institution has digital communication channels with families, the neutrosophic stance analysis of Chapter 7 can be applied to messages from the last three months to detect parental attitudes before launching any formal communication.

Target proposition: "The integration of AI in the classroom at [institution name] will benefit my children's learning."

11.3.4 Pilot Group Selection

- **1–2 lead teachers** (Favourable consensus zone): those who will implement the pilot.
- **1–2 pilot courses** (recommended: one upper secondary and one lower secondary equivalent): sufficient diversity without excessive complexity.

- **Control group** (at least one equivalent course without the system): essential for claiming improvement is due to the intervention.

The Diagnosis produces a single deliverable: a **2–3 page document** that honestly describes the starting point. It is not a document for the Ministry; it is the institution's internal contract with itself.

11.4 Phase 1 — Preparation (Months 3–4)

Three products: trained teachers, working bots, and approved policy. Without all three, the pilot should not begin.

11.4.1 Teacher Training in the 4D Framework (8 Hours)

Table 11.3. Teacher training workshop structure.

Module	Duration	Content	Module product
1	2h	The problem: why classical evaluation does not work with AI. Real cases.	Teacher identifies in their subject which evaluation formats can be delegated to AI
2	2h	The 4D framework + evaluation tensor (simplified, no maths). EcuaEval demo.	Teacher redesigns ONE task with the 4D framework
3	2h	Building EcuaTutor: each teacher configures their own bot in Poe or Claude Projects	Working bot for the teacher's subject
4	2h	Scenario testing + system prompt iteration + institutional AI policy	Feedback on policy draft; pilot commitment

The workshop is not a conference about AI. It is a construction workshop: by the end, each participating teacher has a working bot, a redesigned task, and an operational understanding of the framework.

11.4.2 Building Institutional Bots

EcuaTutor by subject: using the Chapter 10 system prompt, personalised for each subject with the course content corpus. Minimum recommended: 2–3 EcuaTutors for the pilot subjects.

Institutional EcuaEval: using the Chapter 9 prompt (Lab 9.A), configured with 4D criteria specific to the educational level.

11.4.3 The Institutional AI Policy (One Page)

```
AI USE POLICY — [INSTITUTION NAME]
Approved by: [signature] | Date: _____

WHAT WE PERMIT
- Using AI for research, synthesis, and practice (Open Lane)
- Using EcuaTutor as out-of-class study support
- Using AI to create first drafts that the student subsequently reviews

WHAT WE DO NOT PERMIT
- Submitting as one's own a text generated entirely by AI without review
- Using AI during Safe Lane assessments (exams, oral defences)
- Sharing personal student data with non-approved AI platforms

WHAT WE EXPECT
- Students to document their AI use (process portfolio)
- Teachers to review and approve any AI-generated evaluation
- Families to know and understand this policy before the pilot starts

CONTACT FOR QUERIES AND COMPLAINTS
[responsible teacher contact and administrator contact]
```

The policy is communicated to families at an in-person meeting before the pilot begins, not by circular letter.

11.5 Phase 2 — Pilot (Months 5–8)

11.5.1 Week 1: Controlled Launch

Monday: presentation of EcuaTutor to pilot group students (30 minutes in class). Teacher demonstrates live: how to access, what to ask, what not to request. First guided in-class session.

Tuesday–Thursday: first task with process portfolio. Student documents at least one EcuaTutor session using the minimum prompt log: what they asked, what the bot responded, what they decided to do with that response.

Friday: first feedback session. Not about the content of the task: about the *process*. When did they find the bot useful? When frustrating? When did they ignore it?

11.5.2 Weekly Monitoring with the Tensor

From week 2, monitoring uses the Student Evaluation Tensor. The teacher does not need to fill the complete tensor each week: updating one entry per student per week is sufficient — the one corresponding to the most relevant criterion for the current task and the EcuaEval evaluator.

If mean group UMC falls or stagnates for two consecutive weeks, the most common causes are: EcuaTutor corpus outdated relative to class progress, scaffolding threshold miscalibrated, or students struggling with the process portfolio rather than content.

11.5.3 Mid-Pilot Evaluation (Month 7)

Comparative evaluation between the pilot group and the control group using the same baseline instrument.

Metrics that matter:

- Academic: performance on standardised assessments; task completion rate; process portfolio quality (4D rubric)
- Process: group mean UMC; EcuaTutor sessions per student; zone distribution in SET
- Teacher: preparation and grading hours; satisfaction (neutrosophic Likert, 5 items); number of system prompt adjustments

11.5.4 The Three-Alert Rule

The pilot should be paused if three or more of these five signals appear simultaneously:

21. More than 30% of students used EcuaTutor zero weeks after the second.
22. Group mean UMC did not grow in eight pilot weeks.
23. More than 40% of pilot teachers report the system increased their workload.

24. A case of inappropriate AI use not manageable with the approved policy is documented.
25. The administration receives more than five formal family complaints in the same month.

No individual signal is sufficient to pause the process; three simultaneous signals indicate a systemic problem.

11.6 Phase 3 — Scale (Months 9–12)

If the pilot produces positive or neutral results with at least one clear improvement signal, scaling proceeds. If the pilot was clearly negative, document it, share it with the educational community, and wait. An institution that honestly communicates a failed pilot has more credibility than one that conceals it.

11.6.1 Cascade Teacher Scaling

The most efficient scaling mechanism is the cascade: pilot teachers train their peers. Not the administrator, not an external consultant — the teacher who used the system, saw what worked and what did not, and can answer questions that only someone who was in the classroom can answer.

The cascade training follows the same 8-hour workshop structure, with two modifications: Module 1 is replaced by the pilot teacher's testimony; Module 3 becomes a personalisation workshop rather than construction from scratch.

11.6.2 Replication Documentation (20 pages maximum)

At the end of month 12, the institution produces a replication document including: baseline and final results, EcuaTutor and EcuaEval system prompts adjusted for their context, institutional AI policy with pilot modifications, three things that worked / three that did not / three they would do differently, and actual final budget.

11.7 Budgets for Three Contexts

Table 11.4. Itemised budget by context.

Item	Rural public (USD)	Urban public (USD)	Private (USD)
Teacher training (8h workshop × N	0–200	200–500	500–1,500

teachers)			
Connectivity (WiFi or data improvement)	0–500	500–2,000	0 (already exists)
Additional devices	0 (rotation with existing)	0–2,000	0–3,000
AI platforms (bots)	0 (Poe free)	0–200/month	0–500/month
Coordination (teacher time)	0 (internal)	0 (internal)	0–1,000
Evaluation and measurement	0–100	100–300	300–1,000
Family communication	0–50	50–200	200–500
TOTAL ESTIMATE	**0–850**	**850–5,200**	**1,000–7,500**

11.8 Risk Matrix and Mitigations

Table 11.5. Transformation risk matrix.

Risk	Probability	Impact	Mitigation
Active teacher resistance	High	High	Show the system reduces teacher workload; never present it as teacher evaluation
Students using AI to evade learning	High	Medium	EcuaEval + process portfolio + Safe Lane assessments; no perfect solution
Connectivity or device shortage	High (rural)	High	Design for real context; rotation model; do not wait for ideal conditions
Family concerns	Medium	High if unmanaged	In-person meeting

			before start; transparency about what the system does and does not do
Bot produces incorrect content	Medium	Medium	Teacher-curated corpus; session report review; "EcuaTutor proposes, teacher approves" policy
Administrator loses momentum mid-process	Medium	Very high	Written commitments before starting; visible milestones at months 3 and 7; a lead teacher who does not depend solely on the administrator
Single-platform dependency	Low	High	Document system prompts for portability; never store student data only in the platform

The most underestimated risk in practice is administrator momentum loss. Educational transformations that fail usually do so not because the technology did not work. They fail because the administrator who championed them was replaced, changed priorities, or never transferred leadership to teachers. The structural mitigation is to build from the start a community of practice that survives any individual person.

11.9 What Changes When the System Works

Teachers change what they do with their time. Instead of correcting 35 AI-generated identical essays, the teacher reviews 35 session reports that EcuaTutor

generated, each with three specific recommendations already formulated. Correction time falls. Meaningful feedback time rises. Teachers who lived the change describe it consistently: "Before I was correcting. Now I am teaching."

Students change what they do when they do not understand. Instead of copying a classmate's answer or waiting for the next class, the student opens EcuaTutor at 11 pm and works with it until they understand. 24/7 availability of the Socratic tutor does not replace the teacher: it frees the teacher from questions the bot can answer so the teacher can focus on the ones it cannot.

The institutional conversation about assessment changes. An institution that adopts the SET has, for the first time, evidence that learning is a process and not an event. A group's UMC that grew from 0.42 to 0.71 during the semester is a datum that a grade average cannot produce.

References

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. David McKay.

Carroll, J. B. (1963). A model of school learning. Teachers College Record, 64(8), 723–733. https://doi.org/10.1177/016146816306400801

Hattie, J. A. C. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge. https://doi.org/10.4324/9780203887332

Rogers, E. M. (2003). Diffusion of innovations (5th ed.). Free Press.

Smarandache, F. (1998). Neutrosophy: Neutrosophic probability, set, and logic: Analytic synthesis & synthetic analysis. American Research Press.

UNESCO. (2023). Guidance for generative AI in education and research. UNESCO. https://doi.org/10.54675/EWZM9535

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

Chapter 12. Teaching Uncertainty

The most important skill of the twenty-first century is knowing what you do not know

> *"The greatest enemy of knowledge is not ignorance,*
> *it is the illusion of knowledge."*
> *— Stephen Hawking*

12.1 A Nine-Year-Old and the Thesis of Two Books

In October 2024, a third-grade student in Cuenca was working with EcuaTutor on a science assignment about Ecuador's volcanoes. She asked for the exact date of a specific Tungurahua eruption. The tutor responded:

"I am not completely certain of the exact date of that specific eruption. What I can tell you with more confidence is that Tungurahua had significant eruptions between 1999 and 2016. For the exact date you need, I recommend verifying with your teacher or an official source such as the Instituto Geofísico."

The girl paused. Then she typed:

"That's okay. My teacher says the best scientists are the ones who know when they don't know."

She was nine years old. She had just formulated with precision the central thesis of this book — and the one before it.

Figure 12.1. Illustration of the opening scene: a young girl at a simple school desk, looking at a tablet screen. On the screen, a friendly robot character displays a speech bubble: "I'm not completely certain about this. Let's verify with your teacher." The girl has a thought bubble showing her teacher's face with a quote mark. The mood is curious and warm, not technological or cold.

She did not learn this from an epistemology curriculum. She did not learn it from a critical thinking class. She learned it because her teacher said it aloud in front of the class, and because the digital tutor she used for her homework modelled it in every session: declaring uncertainty is not a failure of knowledge. It is the first act of intellectual honesty.

This final chapter argues that that ability — knowing what you do not know, and having the courage to say it — is the most important a student can develop in the AI era. Not the most elegant, not the most philosophical. The most practically valuable. And it can be taught. From age six. Without complicated formulas and without special budgets. With a teacher who has the habit of saying "I don't know" and a digital tutor that does the same.

12.2 Why Uncertainty Is the Skill of the Century

The trajectory of the last few years suggests that, in the domains they master, the most advanced language models are likely to produce correct responses in the large majority of cases — though any precise figure or date would be exactly the kind of confident, unqualified forecast this book teaches readers to distrust. We do not know the rate, and it will vary sharply by domain. What seems robust is the direction, not the number: in well-

trodden domains, AI's baseline competence is rising fast enough to change the pedagogical question radically.

For two hundred years, education was fundamentally about knowledge mastery: knowing more than the textbook, knowing more than the previous student, knowing enough to answer examination questions. The economic and social value of knowledge depended on its scarcity: the doctor was valuable because they knew what the patient did not, the lawyer because they knew the law the client did not, the engineer because they mastered the mathematics the operator lacked.

AI eliminates that scarcity. Not completely, not permanently — but enough that the central pedagogical question is no longer "how much does the student know?" but "what does the student do with what AI does not know?"

Whatever the residual error rate turns out to be, that residual is not a minor leftover — it is the part that matters most. High-risk medical decisions, diagnoses in unprecedented cases, rulings in legally novel situations, policy recommendations in data-scarce or contradictory contexts: that is precisely the territory where AI tends to err with the same confidence with which it is right, and where a professional without training to detect that miscalibrated confidence can cause real harm.

The professionals best prepared for this shift are not those who know more than AI. They are those who can recognise **when AI does not know**.

And that can be taught. From the first cycle of primary school.

12.3 How Students Learn Uncertainty by Developmental Stage

Table 12.1. Pedagogical progression of uncertainty management by developmental stage.

Age	How they understand "I don't know"	Teaching strategy	Mental model to introduce
6–8	"The robot isn't sure" natural curiosity	Tutor says "I'm not sure, let's find out together"	Map ≠ Territory ("the photo is not the real thing")
9–11	"Some things have more than one answer" nuance	Tutor shows two different answers to the same question	Inversion ("what if the opposite were true?")
12–14	"Evidence can be	TIF with simplified	First Principles

	strong or weak" critical thinking	language; traffic light system	("what do we actually KNOW?")
15–17	"Experts disagree and that is information" intellectual maturity	Full TIF; paraconsistency; debate format	Second-order thinking + Circle of Competence
18+	"I can evaluate the quality of any claim" professional competence	Full framework + domain-specific prompts	Full mental model lattice + TIF as meta-model

12.3.1 Ages 6–8: The Robot That Doesn't Know

Children in this range already have intuition about the difference between a person who knows and a person who guesses. What they do not yet have is the vocabulary or framework to apply it to a machine that sounds confident.

The most effective strategy here is repeated exposure to a tutor that declares its uncertainty in a consistent format and without embarrassment: "I am not sure about that. Let's look it up together." Not as a notable exception, but as normal system behaviour.

Children ages 6 to 8 learn by imitation before instruction. If the tutor models uncertainty calmly and treats it as a starting point for exploration rather than a defect, the child incorporates that attitude without needing an explicit epistemology lesson.

12.3.2 Ages 9–11: Two Answers

In this range, students can handle the coexistence of divergent answers without the discomfort it produces at younger ages. This is the moment to explicitly introduce the contrast: "Here are two different answers to the same problem. Which is more reliable? How would you know?"

The Inversion model can be introduced here: "If the opposite were true, what evidence would I find?" Children of 10 can answer this if given the space to try, and it strengthens D3 (Discernment) of the 4D framework before the student knows it by that name.

12.3.3 Ages 12–14: Evidence Has Weight

This is the range where the TIF framework can be introduced by name with simplified notation. Not with the formulas of Chapter 1, but with traffic light language: green (high T, well-supported), yellow (high I, verify before using), red (high F, contrary evidence).

The key question for this range is not "is it correct?" but "how do you know?" Applied systematically to tutor responses, textbook claims, and the student's own assertions, it builds the epistemic habit that this book has as its goal.

12.3.4 Ages 15–17: Experts Disagree

In this range, students can handle the most sophisticated level of uncertainty: expert disagreement as positive data, not as a signal that "nobody knows anything."

The TIF framework in full and the paraconsistent case concept — high T and high F simultaneously, without that being a contradiction to resolve — are appropriate for this range. Chapter 6 of this book can be used directly with 16-year-old students with basic methodological training.

12.3.5 Age 18 and Above: Professional Competence

In higher and postgraduate education, the goal is not teaching uncertainty as a generic skill. It is teaching it as a domain-specific competence: uncertainty in medical diagnosis has a different structure from uncertainty in legal analysis or public policy recommendations.

Each domain has its own neutrosophic zone profile: which types of assertions tend to be in Consensus, which in chronic Ambiguity, which in productive Contradiction. The expert professional not only knows the content of their field; they know the epistemic map of their field.

12.4 Teaching the Brain to Override Itself: A Proposed Curriculum

Chapter 2 established the evolutionary basis for why truth-seeking is hard. This section proposes the curriculum that teaches the brain to do it anyway — not by pretending the defaults do not exist, but by training systematic compensations that become habit.

The curriculum is called **Epistemic Literacy: Truth, Falsity, and Indeterminacy** (*Alfabetización Epistémica*). It is not a subject for university students who read Kant. It is a subject for ten-year-olds, seventeen-year-olds, and fifty-year-old professionals who have never been taught to ask: *what part of this claim is supported, what part contradicts it, and what part do I genuinely not know?*

12.4.1 The Core Claim: Binary Education Is Insufficient

The dominant model of knowledge evaluation in formal education has been binary for centuries:

```
Claim → True or False?
```

This model works well when claims are clear, evidence is unambiguous, and the student's task is to locate the correct answer. It fails catastrophically in the world we now inhabit, where:

- AI generates plausible claims with uniform confidence regardless of epistemic status
- Information environments mix verified facts, contested interpretations, deliberate misinformation, and genuine uncertainty in the same stream
- Professional decisions increasingly require acting on partial, contradictory, or incomplete evidence
- The cost of false certainty (medical errors, policy failures, financial losses) is asymmetrically high The upgrade is minimal but consequential:

```
Claim → What is true? What is false? What is genuinely indeterminate?
```

This is not philosophical relativism. It is a more accurate description of how knowledge actually works. "The Earth is approximately 4.5 billion years old" is in the Consensus zone (T very high, F very low). "Screen time is harmful to adolescent mental health" is in the Ambiguity zone (T moderate, I high, F moderate — genuine scientific debate). "This specific supplement will cure your specific condition" is in the Dissent zone (F high for most specific claims). The skill is knowing which zone each claim belongs to, and why.

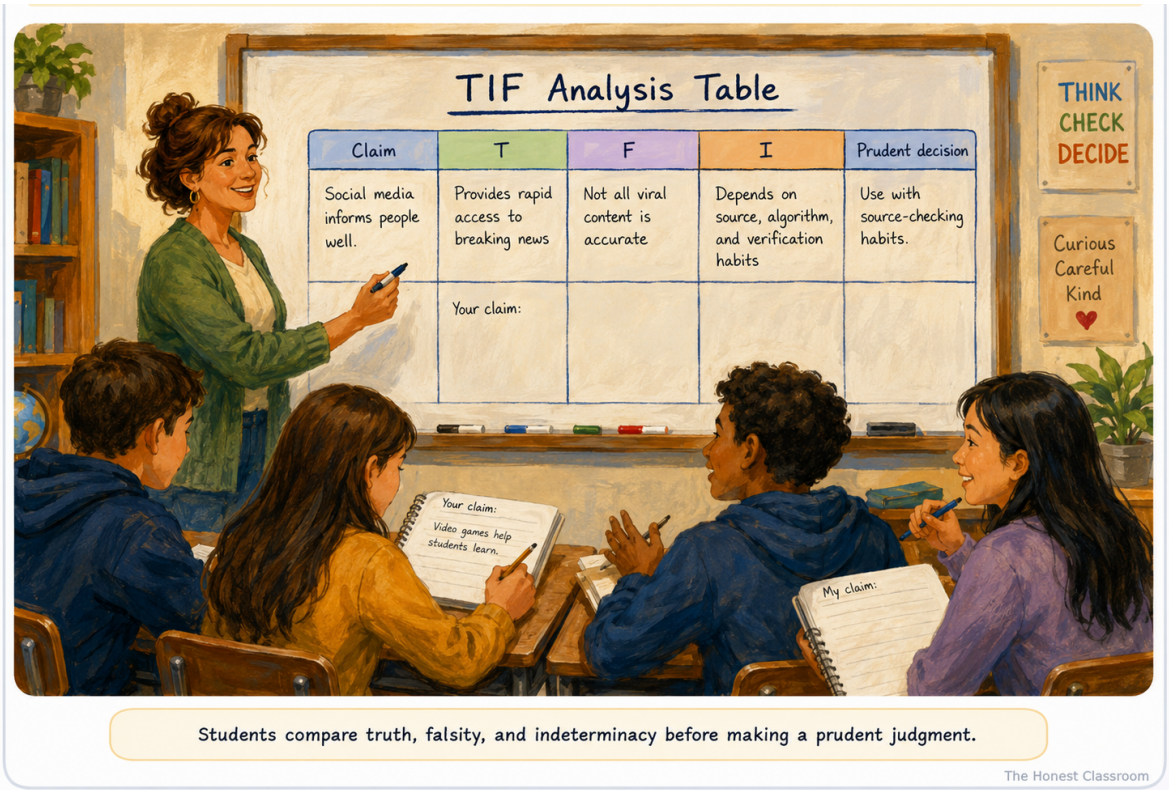

Figure 12.2. A classroom whiteboard showing a four-column "TIF Analysis Table" being used by a teacher. On the board: one row filled in as an example: Claim = "Social media informs people well." | T = "Provides rapid access to breaking news" | F = "Not all viral content is accurate" | I = "Depends on source, algorithm, and verification habits" | Prudent decision = "Use with source-checking habits." Students in the foreground are filling in a second row for their own claim. The atmosphere is engaged, not formal.

12.4.2 The Seven Competencies of Epistemic Literacy

The curriculum develops seven competencies, ordered from foundational to advanced. Each competency corresponds to a specific cognitive default from Chapter 2 and a specific TIF training exercise.

Competency 1 — Slow evaluation (overrides Reactive evaluation)

The student can pause between receiving a claim and evaluating it. They do not respond immediately with agreement or dismissal. They have learned that the first emotional response (like/dislike, agree/disagree based on source) is a signal to activate, not a verdict to adopt.

Training exercise: "Claim Card" — a claim is shown on a card for 30 seconds. Students must write T, I, F values *before* they know who made the claim. Then the source is revealed. Students compare their initial evaluation to their source-adjusted one and name the bias.

Competency 2 — Evidence distinction (overrides Cognitive ease)

The student distinguishes between *feeling* certain and *having evidence* for certainty. They can name the difference between: an anecdote, a correlation, a controlled experiment, a systematic review, an expert opinion, and a model prediction.

Training exercise: "Evidence ladder" — students rank five pieces of evidence about the same claim from weakest to strongest, and explain why each rung is higher than the one below.

Competency 3 — Falsity recognition (overrides Confirmation bias)

The student actively seeks evidence that contradicts their working hypothesis. They understand that finding no contradicting evidence is not proof of truth; it may be proof that they have not looked.

Training exercise: "Steel-man the opposition" — before any argument, the student must articulate the strongest possible case for the opposing view. Only after this are they allowed to argue their own position.

Competency 4 — Indeterminacy tolerance (overrides all four defaults)

The student can say "I don't know yet" without distress. They understand that indeterminacy is not failure; it is an accurate description of the evidence. They can act on partial evidence while flagging what they do not know.

Training exercise: "The I report" — after every major learning unit, students submit a one-paragraph "uncertainty inventory": three things they now understand better (T increased), two things that became more uncertain as they learned more (I increased), and one thing they now know is wrong that they believed before (F revealed).

Competency 5 — Bias detection (applied to self)

The student can name their own cognitive biases in the moment of occurrence. "I notice I'm disbelieving this because it comes from a source I don't trust" is not self-criticism; it is metacognitive competence.

Training exercise: "Bias log" — a running journal where students document three instances per week where they caught themselves applying a cognitive default, and what they did about it.

Competency 6 — Calibrated decision-making under uncertainty The student can make decisions and take actions while maintaining explicit acknowledgement of what they do not know. They understand that acting under uncertainty is not the same as acting recklessly; it is the normal condition of all professional judgment.

Training exercise: "Incomplete case" — students are given a problem with deliberately missing information and must make a recommendation anyway, with explicit acknowledgement of which parts of their reasoning depend on uncertain premises.

Competency 7 — Opinion change without humiliation The student can change a stated position in response to new evidence without losing face. This requires that the classroom culture actively celebrates position changes as evidence of learning, not as defeat.

Training exercise: "The Update" — at the end of a discussion unit, students write: "Before this unit I believed X because Y. Now I believe Z because W. The evidence that changed my mind was..." This is submitted publicly and celebrated.

Competency 8 — Perspectival awareness (advanced, ages 15+)

The student recognises that a different (T, I, F) profile for the same claim does not always mean one person is wrong. It may mean they occupy a different **perspective** — a different position in terms of stakes, history, or values. The student can map the TIF profile of at least two distinct perspectives on a contested claim, identify what each perspective sees that the other misses, and determine whether the disagreement is factual (resolvable with more evidence) or perspectival (stable even with full information).

This competency is grounded in the formal framework of N-alethic Neutrosophic Logic (see Section 12.6.3), which formalises perspective-indexed valuations and the conditions under which inter-perspectival conflict is stable rather than convergent.

Training exercise: "Perspectival TIF Grid" — students evaluate a contested claim from two explicitly assigned perspectives (not their own) plus their own perspective. They complete three rows of the TIF Matrix, one per perspective, and answer: *What sustains each view even in the face of contrary evidence? Is this a disagreement about facts, about stakes, or about values?* The exercise concludes with the student classifying the disagreement: resolvable, partially resolvable, or non-synthetic stable.

12.4.3 The TIF Classroom Analysis Matrix

The central pedagogical tool of the curriculum is the TIF Classroom Matrix — a structured table that replaces the binary "is this true or false?" prompt with a four-column analysis. It can be used at any level from grade 4 onward.

Structure:

Example 1 — Health (grade 6–8):

Example 2 — AI in education (grade 10–university):

Example 3 — Social/political (grade 10–adult):

12.4.4 Why Indeterminacy Is Not Relativism

The most important clarification of the entire curriculum: teaching students to say "I don't know" is not the same as teaching them that all claims are equally valid.

Relativism says: "Everything is a matter of opinion. There is no objective truth."

Epistemic literacy with TIF says: "Some claims have strong T (well-supported), some have strong F (well-refuted), and some are genuinely in the I zone (evidence insufficient to decide). These are different states, and the difference matters for how we act."

The student who has completed this curriculum:

- Knows the difference between *"we don't know yet"* (I, temporary, resolvable with more evidence) and *"it's all subjective"* (a false conclusion that eliminates the possibility of better or worse evidence).
- Can say "this is my best current judgment, not the final truth" without abandoning the claim.
- Understands that some questions are in Zone I for structural reasons (ethical dilemmas, deep values conflicts, irreducibly complex systems) and some are in Zone I for epistemically contingent reasons (we haven't done the study yet).

The clearest test: a student who has learned epistemic literacy should be *more certain* about well-established facts (evolution, the age of the universe, vaccine safety) than a student who has not, because they have learned to recognise when T is genuinely high and evidence is strong. Epistemic literacy produces better calibration — not universal scepticism.

12.4.5 Implementation: From Single Lesson to Full Subject

Minimum viable implementation (one lesson per week, any subject):

Any teacher in any subject can begin with one simple protocol: before any assessment, students must complete one row of the TIF Matrix on the most important claim of the

unit. This takes 10 minutes, produces diagnostic data on student understanding, and begins building the epistemic habit.

Intermediate implementation (embedded across subjects, 1 year):

A school that integrates TIF Matrix analysis into three subjects simultaneously — science, social studies, and language arts — creates cross-domain transfer: the student learns that the three-question protocol applies equally to a scientific claim, a historical interpretation, and a news article.

Full subject implementation (dedicated class, 1 semester):

A standalone "Epistemic Literacy" module of 15-20 sessions covering all seven competencies of Section 12.4.2. The module produces an individual **Epistemic Portfolio**: a collection of TIF analyses across multiple domains, a bias log, and a final reflective essay on how the student's thinking changed during the module.

Table 12.2. Implementation pathway by institutional context and resource level.

12.4.6 The Honest Teacher as the First Lesson

No curriculum document, tool, or framework replaces the most powerful pedagogical intervention available: a teacher who says "I was wrong about that, and here is what changed my mind."

That sentence does four things simultaneously:

26. Models that changing positions is evidence of intellectual honesty, not weakness (trains Competency 7)
27. Makes the process of belief revision visible and learnable (trains Competency 5)
28. Demonstrates that authority and uncertainty are compatible (reduces identity defence bias)
29. Establishes that the classroom is a safe place to be uncertain (reduces strategic silence)

The curriculum provides the tools. The teacher provides the permission. Without both, neither works.

12.5 The Three Things AI Cannot Teach

12.5.1 Motivation

AI can explain fractions. It cannot make a student care about fractions.

Motivation to learn is not a cognitive state; it is an affective-relational state that depends on how the student sees themselves in relation to knowledge and the community of people who value it. A student who does not believe they can learn mathematics does not learn mathematics with a neutrosophic Socratic tutor: they learn to answer the tutor's questions, which is a completely different activity.

What changes a student's relationship with knowledge is not the quality of the explanation. It is the presence of an adult — teacher, family member, mentor — who believes in the student's capacity before the student has evidence to believe in themselves. That has no possible representation in a system prompt.

12.5.2 Emotional Intelligence

AI can detect frustration in student text. It cannot emotionally sustain that student.

"Holding" is the developmental psychology term for something good teachers do naturally: being present with the student's emotional state without trying to resolve it immediately, without dismissing it, without instrumentalising it to produce learning. A student whose parents are separating does not need EcuaTutor to detect "high frustration, switch to basic scaffolding." They need an adult who says "this is hard" and asks nothing in return.

EcuaTutor can have escalation protocols: if it detects significant emotional distress signals, it can generate a notification to the teacher. But the escalation protocol is not the intervention. The intervention is the teacher.

12.5.3 Situated Ethical Judgement

AI can present options. It cannot decide what is right for this student, this community, this moment.

Ethics is not a lookup table. It is the practice of weighing competing values in specific contexts with partial information and uncertain consequences. An AI that "solves ethical problems" is producing the most plausible response in its training corpus, not the most appropriate judgement for the consultee's particular situation.

A student who learns to delegate ethical judgement to AI — "the chatbot told me it was okay" — is not developing ethical judgement. They are developing dependence on an oracle. The honest classroom does not only teach students to evaluate the epistemic quality of AI responses. It teaches that some questions have no answer in any model and require the deliberation of people who know the context, have stakes, and are responsible for the consequences.

12.6 The Honest Teacher as Epistemic Role Model

The girl in Cuenca learned to value "I don't know" because her teacher said it aloud. That is not an anecdotal detail. It is the central pedagogical mechanism of this chapter.

Students do not learn attitudes from curricula. They learn attitudes from the adults in front of them.

A teacher who says "I don't know" in class simultaneously does three things that no curriculum can do alone:

First: demonstrates that not knowing is compatible with authority and competence. The teacher who does not know does not lose stature before students who respect them; they gain it, because they demonstrate that knowledge is an open enterprise.

Second: creates permission for the student not to know. In a classroom where the teacher's uncertainty is normal, the student's uncertainty can be too. It stops being a failure signal and becomes a legitimate position from which to investigate.

Third: models the process of reducing uncertainty. The teacher who says "I don't know, but this is what I would do to find out" teaches more methodology in that sentence than ten classes on the scientific method.

12.6.1 EcuaTutor as Amplifier of the Teacher Model

When EcuaTutor says "I am not completely certain about this in the course materials," it is repeating, in every session with every student, the same message that the honest teacher transmits in their best classroom moments. The difference is scale: a teacher may have four honest "I don't know" moments per week in front of their students. EcuaTutor has that behaviour encoded in its Rule 4 and executes it in every session, with every student, without exception.

It does not replace the teacher. It amplifies the best of the teacher at a frequency the teacher alone cannot sustain.

The necessary condition is that the teacher and the bot transmit the same message. A teacher who punishes uncertainty in the classroom and an EcuaTutor that models "I don't know" produce a contradictory system. Coherence between teacher behaviour and bot behaviour is not an implementation detail; it is the system's operating condition.

12.7 The Latin American Contribution: When "I Don't Know" Already Has a Name

12.7.1 Yanantin: The Complementarity of Opposites

Andean thought has the concept of *yanantin* for the relationship between opposites that do not exclude but complement and sustain each other: day and night, up and down, masculine and feminine. In Western modern thought, two opposites generate a contradiction that must be resolved (thesis, antithesis, synthesis). In Andean thought, two opposites generate a tension to be inhabited as a condition of the world's balance.

Yanantin is not relativism. It does not say everything goes. It says reality is richer than any pair of excluding categories, and wisdom consists in learning to operate productively in the tension between them.

For the pedagogy of uncertainty, this has a direct translation: when the student finds that two well-supported assertions contradict each other — when T is high and F is high simultaneously, the paraconsistent case of Chapter 6 — the correct response is not "one of the two must be wrong." The correct response may be: "both are true under different conditions, and my work is to understand which conditions activate each."

That is not confusion. It is epistemic maturity. And *yanantin* is its oldest name.

12.7.2 Probable Conscience: Certainty Proportional to Gravity

The Latin American Scholastic tradition developed the concept of *probable conscience* as a guide for action under moral uncertainty: when there is no absolute certainty about what is correct, action is calibrated in proportion to the gravity of possible consequences. More serious matters require greater certainty before acting.

Translated to the neutrosophic framework: not all uncertainty requires the same response. A high I about the exact date of a volcanic eruption in a school assignment does not require the same action as a high I about the correct dose of a medication in a clinical context. The proportionality principle — act more carefully when consequences are more serious — is the practical complement to TIF.

12.7.3 N-alethic Epistemics: When the Disagreement IS the Knowledge

Yanantin gives the classroom a philosophical foundation. *Probable conscience* gives it a decision rule. A third concept — which we call N-alethic neutrosophic reasoning — gives it a precise vocabulary for the most sophisticated epistemic situation students encounter: when two well-calibrated observers look at the same event and produce genuinely different (T, I, F) valuations, not because one is wrong, but because they occupy different **perspectives**.

N-alethic logic is not relativism any more than yanantin is. It asserts three things that can be taught from age fifteen onward:

First: valuations are perspectival. The evaluation differs by perspective . A student from a family directly affected by an environmental policy will produce a different TIF evaluation of "this industry generates economic development" than a student whose family owns land adjacent to the affected area. Both TIF profiles are real. Neither is simply wrong. The disagreement is data about the social distribution of benefit and harm.

Second: not all disagreements converge. Classical education treats conflict as a problem to be resolved: debate until the better argument wins. N-alethic epistemics introduces a more honest category — the **non-synthetic stable conflict** — where two perspectives hold genuinely incompatible (T, I, F) profiles and no amount of evidence forces convergence, because the disagreement reflects different values or different stakes, not different facts. Recognising this category is not cynicism. It is precision. It stops students from expecting resolution where none is available, and directs them instead toward productive coexistence with the tension.

Third: force sustains perspective. Each perspective is maintained not only by evidence but by what Nietzsche called *will to interpret* — the interests, histories, and commitments that make certain framings feel obvious. In the classroom, this means: when two students produce radically different TIF valuations of the same claim, the appropriate question is not only "which evidence is better?" but also "what makes each reading feel natural to its holder?" This is a question about power and position, not only about epistemology.

Classroom application (ages 15–18+):

The exercise is called the **Perspectival TIF Grid**. Students evaluate a contested claim — an environmental policy, a public health recommendation, a historical judgement — from two explicitly assigned perspectives, not their own:

Perspective	T	I	F	Force sustaining this view
: *stakeholder directly affected*	Evidence that supports the claim from lived impact	What this group cannot yet know or verify	Evidence of harm, exclusion, or cost	Stakes, history, vulnerability, or dependence
: *technical*	Evidence that	Limits of the	Evidence that	Disciplinary

expert outside the community	supports the claim from models, data, or precedent	model, dataset, or method	contradicts the technical conclusion	training, institutional role, or methodological commitments
My own perspective	Evidence I find convincing after hearing both positions	Questions I still cannot answer responsibly	Counter-evidence I must keep visible	My values, interests, prior beliefs, or social position

After completing the grid, students answer: *What would it take for these perspectives to converge? Is convergence possible, or is this a productive non-synthetic tension?*

This exercise does not produce a "correct answer." It produces epistemic precision about where the disagreement actually lives: in the evidence, in the stakes, or in the values. That distinction is one of the most practically valuable a student can develop.

12.8 The Vision for 2030: The Classroom That Becomes Honest

Imagine a classroom in 2030 where this book worked.

Not a perfect classroom. An honest classroom.

The AI tutor does what it does well: delivers personalised content, one-to-one, mastery-based, available at eleven at night when the student has a question about tomorrow's assignment. Declares its uncertainty in a consistent format. Generates the evaluation tensor in real time. Produces reports the teacher can read in ten minutes.

The teacher does what only they can do: knows their students by name and story. Knows which one is going through grief, which one has difficulties at home, which one needs someone to believe in them before they believe in themselves. Teaches the moments that fit in no prompt: the instant when a difficult concept clicks, the conversation about ethics that has no right answer, the debate where two well-argued positions meet and neither fully prevails.

The student has internalised three questions that were not in the official curriculum but structure everything else: What supports this claim? What do we not yet know? What contradicts it? They apply them to what the tutor tells them, to what the teacher tells them, to what the news tells them, to what they tell themselves when reasoning about a difficult problem.

The institution has data it did not have before: not the group's grade average, but each student's tensor profile, the UMC trajectory during the semester, the neutrosophic zone distribution by criterion and teacher. Those data do not replace the teacher's judgement. They inform it.

Figure 12.3. Vision illustration: a classroom of the future that looks warm and human, not technological or futuristic. A teacher stands at the front, genuinely engaged with a small group of students discussing something. On individual tablets around the room, EcuaTutor sessions are running silently. One student's tablet shows a TIF assessment report. The mood is of active, curious inquiry. Natural light through windows. No dramatic technology — just humans and tools working together.

And in that classroom, the most natural thing in the world is for a ten-year-old to look at a confident screen and say: "Wait. How sure are you about that?"

Not because they were taught to distrust technology. But because they were taught to think.

12.9 What This Book Did Not Solve

The access gap. The model works better with stable connectivity, available devices, and teachers with time to train. Many of the classrooms that most need this framework are exactly those with the least of those resources.

Deep institutional resistance. There are institutions where pedagogical change is structurally impossible while system incentives point in another direction: standardised assessments that measure memory, accreditation systems that reward process over

learning outcomes, organisational cultures where the experimenting teacher is penalised for the failure.

The speed of AI change. The models that exist when this chapter is written are not those that will exist when it is read. The principles — manage uncertainty, evaluate the process, teach discernment — are more stable than the specific tools. But some tools this book recommends may be obsolete before the semester ends.

The judgement problem. The TIF framework, the SET, the UMC: all are instruments that improve the information available for the teacher's judgement. They do not replace that judgement. A teacher who applies the tensor mechanically without reading it in light of what they know about the specific student is doing exactly what the book criticises: delegating thinking to a system that does not know the context.

These limitations do not invalidate the framework. They complete it. A book that does not declare its own I gaps has a coherence problem with its central thesis.

12.10 The Last Question

The girl in Cuenca wrote: "My teacher says the best scientists are the ones who know when they don't know."

That sentence has two parts worth separating.

The first part — "the best scientists" — could have been said by a teacher in 1950. It is not new. Intellectual honesty as an epistemic virtue has a long history.

The second part — "when they don't know" — *is* new in the context of 2024. Before, the question of when you do not know was primarily about yourself: when your memory fails, when your reasoning has gaps, when your evidence is insufficient. Now it is also about the screen: when the system that produces text with limitless confidence is generating a fabrication that looks identical to a truth.

Teaching that to a nine-year-old does not require neutrosophy or tensors. It requires a teacher who says "I don't know" aloud and a tutor that does the same. The rest of this book is the formal scaffolding to sustain that simple act at institutional scale.

The honest classroom is not the classroom without AI.

It is the classroom where everyone — teacher, student, and machine — has the habit of saying "I don't know" before being asked, and the capacity to keep working from there.

That is what we teach when we teach uncertainty.

And it is the most important thing we can teach.

References

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. David McKay.

Dweck, C. S. (2006). Mindset: The new psychology of success. Random House.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. Pantheon Books.

Hattie, J. A. C. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge. https://doi.org/10.4324/9780203887332

Hawking, S. W. (1988). A brief history of time: From the Big Bang to black holes. Bantam Books.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Leyva-Vázquez, M. Y., & Smarandache, F. (2026). The third answer: Why AI doesn't know what it doesn't know—and how ancient logic can fix it: A framework for thinking clearly in the age of confident machines. Neutrosophic Science International Association Publishing House.

Novak, J. D. (1998). Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations. Lawrence Erlbaum Associates.

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. Nature Human Behaviour, 3(2), 173–182. https://doi.org/10.1038/s41562-018-0506-1

Ortiz Rescaniere, A. (1973). De Adaneva a Inkarrí: Una visión indígena del Perú. Retablo de Papel.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. Theory Into Practice, 41(4), 219–225. https://doi.org/10.1207/S15430421TIP4104_3

Plato. (2002). Meno (G. M. A. Grube, Trans.). In J. M. Cooper & D. S. Hutchinson (Eds.), Plato: Complete works (pp. 870–897). Hackett Publishing.

Smarandache, F. (1998). Neutrosophy: Neutrosophic probability, set, and logic: Analytic synthesis & synthetic analysis. American Research Press.

Stanovich, K. E. (2011). Rationality and the reflective mind. Oxford University Press.

Twenge, J. M. (2017). iGen: Why today's super-connected kids are growing up less rebellious, more tolerant, less happy—and completely unprepared for adulthood. Atria Books.

Appendix A — Operational Templates

Ready to use. Copy, adapt to context, implement.

A.0 TIF Claim Analysis Matrix (classroom tool)

For use from Grade 4 through postgraduate. One row per claim to analyse.

Full version (print on A4 landscape)

```
┌──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────┐
│ TIF ANALYSIS MATRIX │
│ "Every claim has three dimensions. Always ask all three." │
├──────────────────────┬──────────────────────┬──────────────────────┬──────────────────────┬────────┤
│ CLAIM │ T — Evidence that │ F — Evidence that │ I — What we │PRUDENT │
│ (what is claimed) │ SUPPORTS │ CONTRADICTS │ DON'T YET KNOW │DECISION│
│ │ (+) │ (-) │ (?) │ │
├──────────────────────┼──────────────────────┼──────────────────────┼──────────────────────┼────────┤
│ │ │ │ │ │
│ │ │ │ │ │
├──────────────────────┼──────────────────────┼──────────────────────┼──────────────────────┼────────┤
│ │ │ │ │ │
│ │ │ │ │ │
├──────────────────────┼──────────────────────┼──────────────────────┼──────────────────────┼────────┤
│ │ │ │ │ │
│ │ │ │ │ │
└──────────────────────┴──────────────────────┴──────────────────────┴──────────────────────┴────────┘

REMEMBER:
• High I does NOT mean "anything goes" → it means "I need more evidence before deciding"
• You can act even when I is high → act proportionally to the evidence you DO have
• Changing your mind when evidence changes is NOT weakness → it is epistemic competence
```

Classroom usage guide (5 steps)

```
Step 1 → Present the claim WITHOUT revealing the source yet
Step 2 → Students fill in T, F, I individually (5 minutes)
Step 3 → Share in pairs; identify differences in assessments
```

```
Step 4 → Reveal the source; students update if their evaluation changes
Step 5 → Class discusses: what bias changed when we knew who said it?
```

Signs of student progress

- Can articulate I without embarrassment: "I don't know this yet"
- Can name their own bias: "I rejected it because I don't trust that source, not because the evidence is bad"
- Can change position without feeling defeated: "The evidence they presented raised my T"
- Can act despite high I: "I don't have total certainty but the most prudent decision given what I know is..."

A.1 TIF Quick Reference Card for Teachers (one printable page)

```
┌──────────────────────────────────────────────────────────────────┐
│ NEUTROSOPHIC COMPASS — QUICK REFERENCE │
│ The Honest Classroom, 2026 │
├──────────────────────────────────────────────────────────────────┤
│ │
│ THE TRIPLE: (T, I, F) — each value between 0 and 1 │
│ │
│ T = What SUPPORTS this claim? (evidence in favour) │
│ I = What do we NOT YET KNOW? (genuine uncertainty) │
│ F = What CONTRADICTS it? (evidence against) │
│ │
│ IMPORTANT: T + I + F does not have to equal 1 │
│ If T + F > 1 → productive ambivalence (paraconsistent) │
│ │
├──────────────────────────────────────────────────────────────────┤
│ THE FOUR ZONES │
│ │
│ CONSENSUS T > 0.6, F < 0.3, I < 0.3 │
│ Well-supported. Use with confidence. │
│ │
│ AMBIGUITY I > 0.5 (any T, F) │
│ Insufficient information. Verify before using. │
│ │
│ DISSENT F > 0.6, T < 0.3, I < 0.3 │
│ Evidence against. Do not use without new evidence. │
│ │
│ CONTRADICTION T > 0.5 AND F > 0.5 simultaneously │
│ Both directions have support. Investigate conditions. │
```

```
| |
├──────────────────────────────────────────────────────────────┤
| THE TRAFFIC LIGHT FOR STUDENTS |
| |
|   GREEN: Use AI freely (research, practice) |
|   YELLOW: Use with teacher guidance (analysis, drafts) |
|   RED: No AI (exams, oral defences, demonstrations) |
| |
├──────────────────────────────────────────────────────────────┤
| THE 4D EVALUATION CRITERIA |
| |
| D1 Delegation Did the student use AI appropriately? |
| D2 Description Did the student document their process? |
| D3 Discernment Did the student critically evaluate AI? |
| D4 Diligence Did the student verify sources? |
| |
| UMC (Uncertainty Management Coefficient): |
| Basic 0.30–0.55 | Intermediate 0.56–0.75 | Advanced 0.76–1.00|
| |
└──────────────────────────────────────────────────────────────┘
```

A.2 4D Rubric for a Single Task (one-page version)

```
4D EVALUATION RUBRIC
Subject: _____ Task: _____ Student: _____

INSTRUCTIONS: Assign a TIF triple (T, I, F) to each dimension.
T = evidence that demonstrates the criterion (0.00 to 1.00)
I = evidence that suggests but does not confirm (0.00 to 1.00)
F = evidence that contradicts or absence of criterion (0.00 to 1.00)

┌──────────────────┬───────┬───────┬───────┬──────────────────────┐
| Dimension | T | I | F | Note |
├──────────────────┼───────┼───────┼───────┼──────────────────────┤
| D1 Delegation | | | | |
| (used AI well?) | | | | |
├──────────────────┼───────┼───────┼───────┼──────────────────────┤
| D2 Description | | | | |
| (documented?) | | | | |
├──────────────────┼───────┼───────┼───────┼──────────────────────┤
| D3 Discernment | | | | |
| (evaluated | | | | |
| AI critically?) | | | | |
├──────────────────┼───────┼───────┼───────┼──────────────────────┤
| D4 Diligence | | | | |
| (verified?) | | | | |
```

```
└──────────────────┴────────┴────────┴────────┴──────────────────────┘

UMC estimate: ______ Level:   Basic   Intermediate   Advanced

Main feedback:
_____________________________________________________________________

Closing question for the student:
_____________________________________________________________________

Evaluator: _________________ Date: _________________
  Self-assessment   Peer   EcuaEval (review before delivering)   Teacher
```

A.3 The Four Cognitive Defaults and Their Classroom Corrections (reference card)

(from Chapter 2)

```
┌──────────────────────────────────────────────────────────────────┐
│ THE FOUR EVOLUTIONARY DEFAULTS — QUICK REFERENCE                 │
├──────────────────┬──────────────────────┬────────────────────────┤
│ Default          │ TIF distortion       │ Classroom correction   │
├──────────────────┼──────────────────────┼────────────────────────┤
│ Confirmation     │ F under-weighted;    │ "Name 2 pieces of      │
│ bias             │ T inflated for       │ evidence that would    │
│                  │ confirming info      │ lower your T"          │
├──────────────────┼──────────────────────┼────────────────────────┤
│ Identity         │ I inflated as        │ Anonymous TIF before   │
│ defence          │ social safety        │ group discussion       │
├──────────────────┼──────────────────────┼────────────────────────┤
│ Cognitive        │ I under-reported     │ Mandatory verification │
│ ease             │ (false certainty     │ step before accepting  │
│                  │ from fluent text)    │ any AI claim as T      │
├──────────────────┼──────────────────────┼────────────────────────┤
│ Reactive         │ F inflated by        │ Source-blind           │
│ evaluation       │ source affect,       │ evaluation: evaluate   │
│                  │ not content          │ claim before           │
│                  │                      │ knowing source         │
└──────────────────┴──────────────────────┴────────────────────────┘
```

Appendix B — Policy Brief

Five Principles for an AI-Honest Educational Institution

Addressed to: ministers, deans, rectors, principals, and heads of teacher training institutes.

Length: four pages. Derived from this book's institutional and policy chapters.

> *"An educational policy that denies indeterminacy reproduces it.*
> *One that represents it explicitly can begin to act on it."*

The problem in one paragraph

Large language models produce confident output regardless of whether the underlying knowledge is strong, weak, contested, or fabricated. Educational institutions have responded with one of three inadequate policies: ban AI (unenforceable), embrace AI uncritically (epistemically irresponsible), or say nothing and let each teacher decide alone. The five principles below offer a fourth path: design institutional practice around the explicit representation of indeterminacy, making the institution honest about what it knows, what it does not know, and what tools it deploys.

The Five Principles

Principle 1 — Institutional Declaration of Indeterminacy

What: The institution publishes, annually, the areas in which its own evidence is insufficient to settle its own decisions.

Why: Institutions that project confidence on all fronts produce internal reports indistinguishable from external marketing. An annual declaration of indeterminacy is not a confession of failure — it is the foundation of a research agenda and the precondition for honest self-evaluation.

How: A two-page public document listing 6–12 open institutional questions, classified by zone (Clarity / Ambiguity / Contradiction / Ignorance) with the investigation that would move each to a better-supported zone.

Who does it: Institutional leadership, published annually before the academic calendar begins.

Principle 2 — Mandatory Audit of AI Tools

What: Any AI system incorporated into academic or administrative operations is audited, by an independent body, for its capacity to declare "I do not know" and for its failure modes under adversarial conditions.

Why: Vendors promise accuracy they cannot substantiate. Institutions acquire tools they cannot audit. When those tools make consequential decisions about students (grading, admissions, plagiarism detection), the institution bears legal and ethical responsibility for errors it did not detect.

How: Three-level audit — unit testing (does the tool declare uncertainty?), behavioural testing (does it hallucinate under normal use?), adversarial testing (can it be prompted into harmful outputs?). Protocol derived from Chapter 3 of this book. Audit is a condition of procurement, written into vendor contracts.

Who does it: An internal or external body independent of the procurement decision. References: EU AI Act (2024), NIST AI RMF (2023), UNESCO Recommendation on AI Ethics (2021).

Principle 3 — Rubrics with Explicit Indeterminacy Component

What: Institutional evaluations of students, teachers, and programmes include an explicit component for what is unresolved — the I dimension of the TIF framework.

Why: A rubric that scores "understanding of the material" without scoring "honest identification of what remains unresolved" rewards premature confidence.

A teacher evaluation that scores clarity without scoring epistemic honesty rewards the same vice at the institutional level.

How: For high-stakes evaluations (doctoral defences, programme accreditations, teacher promotion), require that at least one rubric dimension explicitly evaluates the quality of uncertainty management (I-dimension), not only the quality of correct knowledge (T-dimension). Lower-stakes evaluations may use simpler instruments. See the 4D rubric and SET framework in Chapter 9.

Who does it: Department chairs and evaluation committees, supported by faculty development.

Principle 4 — Prohibition of Agnotological Practice

What: The strategic invocation of indeterminacy to obstruct evidence-based decisions is documented, attributed, and sanctioned over time.

Why: Not all uncertainty is genuine. Proctor and Schiebinger (2008)

documented how industries manufactured uncertainty to delay action on settled questions. Educational institutions do this too: "ongoing controversy" is invoked to delay curriculum updates, technology governance, and equity interventions on questions the evidence has settled.

How: Maintain a documented record of indeterminacy claims made in institutional decisions. When a pattern emerges of "controversy" claims in the same topic area, trigger an independent epistemic audit of that area.

The sanction is reputational, not punitive: attribution of the pattern to the relevant decision-makers, visible to the institution's peers and accreditation body.

Who does it: Institutional governance bodies, with faculty senate oversight.

Principle 5 — Continuous Teacher Formation in Epistemic Literacy

What: Annual, mandatory, focused specifically on the distinction between genuine and strategic indeterminacy, and on the practical tools for teaching epistemic calibration.

Why: The four preceding principles rest on the competence of those who implement them. Competence in recognising and representing indeterminacy is not the default competence of educators trained before 2020. Without formation, the policy is aspiration.

How: 15–20 hours per year, structured around:

- The TIF framework applied to course material (Chapter 6 methodology)
- EcuaTutor design and session report interpretation (Chapter 10)
- The 4D evaluation framework and UMC interpretation (Chapter 9)
- Strategic vs. genuine indeterminacy (Chapter 2)

Assessment of formation is a condition of continued teaching.

Who does it: Teacher development offices, with faculty input on domain-specific application. Curriculum provided in the companion repository (github.com/mleyvaz/honest-classroom).

What These Principles Do Not Solve

Adopting the five principles does not address: equity of access to AI tools across socioeconomic strata, linguistic and cultural adaptation of tools built primarily in English, environmental cost of AI deployment, intellectual-property implications of AI-generated student work, or data governance for AI systems that learn from student interactions. Each is a major policy field that the five principles are compatible with but do not substitute for.

The principles solve one problem: the epistemic vacuum created by institutions that deploy confident tools without building the capacity to evaluate their confidence. That problem is solvable now, with the tools in this book.

Implementation Roadmap

Timeframe	Action	Who
Month 1–3	Publish first Annual Declaration of Indeterminacy	Leadership
Month 1–6	Establish AI audit function (internal or contracted)	Governance
Semester 1	Pilot 4D rubric in 3–5 high-stakes evaluations	Pilot faculty
Year 1	First teacher formation cohort (15–20 hours)	Faculty development
Year 2	Audit all AI tools in active use; retire non-compliant	Procurement
Year 2–3	Extend 4D rubric and I-component to all high-stakes	All departments
Ongoing	Annual Declaration + formation + audit cycle	Institutional calendar

Estimated cost for a university of 500–2,000 students:

- Year 1: $8,000–$25,000 (primarily formation programme and audit)
- Ongoing: $3,000–$8,000/year (annual cycle)

These estimates assume use of free-tier AI platforms (Poe, Claude Projects)

and open-source code from the companion repository. Premium platforms increase costs proportionally.

Contact

Maikel Yelandi Leyva-Vázquez · mleyvaz@gmail.com

Universidad Bolivariana del Ecuador · Durán, Ecuador Repository: github.com/mleyvaz/honest-classroom

This brief may be reproduced and adapted freely under CC BY 4.0.

Appendix C — Two Complementary Taxonomies of Indeterminacy: D and I

Two Complementary Taxonomies of Indeterminacy: D and I

Why this chapter is not in the main text

Appendix C proposes two formal taxonomies — a Developmental taxonomy (D –D , describing how students learn to inhabit uncertainty) and an Ontological taxonomy (I –I , classifying the kinds of not-knowing a student faces) — as complements to Bloom's cognitive taxonomy.

The argument is sound and the contribution is original. It does not belong in this book for one editorial reason: it arrives after the book's practical arc is complete (Chapters 6–8) and before the book's emotional close (Chapter 12), and at that point it slows the reader without adding to the tools they need.

Connection to the main text

Readers who want the formal underpinnings of the UMC levels (Chapter 9 §9.7.3)

and the developmental progression table (Chapter 12 §12.3, Table 12.1) will find them in that paper. Specifically:

- The UMC levels Basic / Intermediate / Advanced correspond to D –D of the Developmental taxonomy.
- The neutrosophic **zones** (Consensus / Ambiguity / Contradiction / Dissent)

correspond to I –I of the Ontological taxonomy.

- The full taxonomic argument, with comparisons to Bloom-Anderson-Krathwohl and to Knight-Keynes risk/uncertainty distinction, is in the working paper.

Summary of the argument (for the curious reader)

The gap in Bloom's taxonomy:

Bloom (1956), revised by Anderson and Krathwohl (2001), organises cognitive objectives from Remember to Create. All six levels share an assumption:

the object of cognition is knowable. What the student does is what they do with what they know. The taxonomy has no level for what the student does when what they are working with is not fully knowable.

The Developmental taxonomy (D — five levels):

Level	Description	Classroom observable
D Basic	Recognises that uncertainty exists	Student says "I don't know"
D Aware	Can describe the source of the uncertainty	"I don't know because the evidence is contradictory"
D Analytical	Can quantify and classify uncertainty (TIF)	Applies the zone system to a claim
D Strategic	Acts differently based on uncertainty type	Different decisions for Ambiguity vs Contradiction zones
D Generative	Produces knowledge structures that make uncertainty productive	Designs assessments that reward I-reduction

The Ontological taxonomy (I — five types):

Level	Type of not-knowing	Example	TIF profile
I Ignorance	No information available	Topic not studied	Low T, High I, Low F
I Ambiguity	Information exists but is unclear	Contested definition	Low T, High I, Low F
I Contradiction	Two well-supported positions conflict	Scientific debate	High T, Medium I, High F
I Complexity	System too complex to predict	Social policy outcomes	Medium T, High I, Medium F
I Fundamental	Unknowable in principle	Ethical dilemmas	Variable

Appendix D — Annotated Bibliography: 19 Key Sources

For readers who want to go deeper. Each entry explains why the source matters for this book.

1. Smarandache, F. (1998). Neutrosophy: Neutrosophic probability, set, and logic: Analytic synthesis & synthetic analysis. American Research Press.

The primary source for neutrosophic logic. It defines the triplet (T, I, F) and its fundamental mathematical properties. It is the essential starting point for any work in the field. The text is freely available through the author's website.

2. Wang, H., Smarandache, F., Zhang, Y. Q., & Sunderraman, R. (2010). Single valued neutrosophic sets. Multispace and Multistructure, 4, 410–413.

This article introduces Single-Valued Neutrosophic Sets (SVNS), the operational version of neutrosophy used throughout this book. The distinction between SVNS and general neutrosophy is technical but important for implementation: SVNS works with values in $[0,1]^3$, while general neutrosophic logic allows broader formal ranges.

3. Leyva-Vázquez, M. Y., & Smarandache, F. (2026). The third answer: Why AI doesn't know what it doesn't know—and how ancient logic can fix it: A framework for thinking clearly in the age of confident machines. Neutrosophic Science International Association Publishing House.

The previous book by the same authors. It develops the TIF framework as an instrument for auditing the outputs of language models. The Honest Classroom is its pedagogical application; the two books are complementary.

4. Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. David McKay.

The most influential taxonomy in curriculum design. This book takes Bloom's taxonomy as a starting point and extends it with dimensions for managing uncertainty—dimensions that Bloom did not need to formalize in 1956.

5. Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

This book defines the Zone of Proximal Development (ZPD), a central concept for the design of EcuaTutor. A tutor that does not operate within the learner's ZPD—either too easy or too difficult—does not teach effectively.

6. Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry, 17(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

The foundational article on scaffolding. It defines the core functions of tutoring and the principle of fading. EcuaTutor's rule of never giving the answer directly implements this principle.

7. Hattie, J. A. C. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge. https://doi.org/10.4324/9780203887332

A major synthesis of more than 800 meta-analyses on factors that influence learning. Its central conclusion is that feedback has one of the strongest effects on achievement. It informs the design of EcuaTutor's session report: feedback must be specific and actionable, not generic.

8. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

One of the most cited systematic reviews on hallucination in natural language generation. It classifies hallucination types, including factuality, faithfulness, attribution, and coherence. It provides the foundation for Chapter 1's argument about why language models do not structurally "know that they do not know."

9. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

This article documents the systematic bias of AI-text detectors against non-native English writers. It is key evidence for the book's argument that detectors are not a sufficient solution to the problem of assessment in the age of generative AI.

10. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

A foundational review of sentiment analysis. It defines the standard framework that this book extends through the neutrosophic triplet.

11. Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). Stance and sentiment in tweets. ACM Transactions on Internet Technology, 17(3), Article 26. https://doi.org/10.1145/3003433

This article defines stance detection as an NLP task distinct from sentiment analysis. It is central to the book's treatment of educational discourse analysis: the distinction between emotion and position is essential.

12. Proctor, R. N., & Schiebinger, L. (Eds.). (2008). Agnotology: The making and unmaking of ignorance. Stanford University Press.

This edited volume defines agnotology as the study of deliberately produced ignorance. It is relevant to the chapters on TIF analysis because a high I value may reflect honest uncertainty, but it may also reflect strategic ambiguity or manufactured doubt.

13. Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 22(140), 1–55.

The original article introducing the Likert scale. It remains surprisingly accessible for a text published in 1932. Reading it makes clear that Likert himself was aware of some of the measurement problems later associated with attitude scales, including ambiguity around intermediate responses.

14. Ye, J. (2014). A multicriteria decision-making method using aggregation operators for simplified neutrosophic sets. Journal of Intelligent & Fuzzy Systems, 26(5), 2459–2466. https://doi.org/10.3233/IFS-130916

This article defines aggregation operators for simplified neutrosophic sets, including widely used weighted aggregation procedures in neutrosophic multicriteria decision-making. The tensor-contraction operator developed in this book recovers these aggregation procedures as special cases.

15. Carroll, J. B. (1963). A model of school learning. Teachers College Record, 64(8), 723–733. https://doi.org/10.1177/016146816306400801

A foundational model of school learning prior to the later development of mastery learning. It establishes the principle that learning time may vary across students while achievement standards can remain constant. This informs EcuaTutor's design: the tutor should not advance until the learner's TIF profile indicates sufficient understanding.

16. Dweck, C. S. (2006). Mindset: The new psychology of success. Random House.

This book defines the distinction between a fixed mindset and a growth mindset. EcuaTutor implements growth-mindset language: error is treated as data, not as a verdict on the learner.

17. Rogers, E. M. (2003). Diffusion of innovations (5th ed.). Free Press.

The standard theory of how innovations are adopted in social systems. It is relevant to the institutional implementation chapter: educational transformation does not follow a linear schedule but an adoption curve, in which early adopters must be identified, supported, and connected.

18. UNESCO. (2023). Guidance for generative AI in education and research. UNESCO. https://doi.org/10.54675/EWZM9535

One of the most widely cited international policy documents on generative AI in education. It is useful for institutional implementation because it gives policy legitimacy to the need for governance, transparency, teacher training, and responsible AI use.

19. Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

This article defines metacognition as knowledge about one's own knowledge and control over one's own cognitive processes. The Cognitive Governance Index (CGI) proposed in this book can be understood as a neutrosophic operationalization of metacognition: it measures how well students manage their own uncertainty.

THE HONEST CLASSROOM

Neutrosophic Logic in AI-Driven Education

AI can answer with confidence even when the evidence is incomplete, contradictory, or still unknown. The Honest Classroom proposes a practical educational response: teach students and teachers to distinguish what is true, what is false, and what remains indeterminate.

INSIDE THE BOOK

- Neutrosophic tools for surveys, rubrics, tutoring, and student evaluation
- TIF models for truth, indeterminacy, and falsity
- EcuaTutor and EcuaEval as practical classroom architectures
- A humane approach to academic integrity in AI-assisted writing
- Ready-made templates and implementation guidance for institutions

> **The goal is not to make students distrust AI. The goal is to teach them how to verify it.**

NSIA PUBLISHING

Education · Logic · AI

ISBN 978-1-972502-28-0

9 781972 502280